Research Report

Statistical and Measurement Properties of Features Used in Essay Assessment

Shelby J. Haberman

Research & Development

June 2004
RR-04-21

Statistical and Measurement Properties of Features Used in Essay Assessment

Shelby J. Haberman

ETS, Princeton, NJ

June 2004

Abstract

Statistical and measurement properties are examined for features used in essay assessment to determine the generalizability of the features across populations, prompts, and individuals. Data are employed from TOEFL® and GMAT® examinations and from writing for Criterion℠.

Key words: Essay assessment, human holistic scores, e-rater®

Acknowledgements

Most features used in this report were generated through the assistance of Jill Burstein and Slava Andreyev. The author thanks Neil Dorans, Sandip Sinharay, and Yigal Attali for helpful comments and assistance in interpretation of results.

Introduction

Numerical features of essays are commonly employed in their assessment. These features can include human holistic scores, word counts, error rates, and many other possible numerical measures. Common assessments may be based on human holistic scores alone, as is the case in the Analytical Writing Measure of the GRE®, the writing section of the TOEFL®, the writing section of the SAT® II Writing Test, and the essay section of the PPST® Writing Test. The Analytical Writing Assessment of the GMAT®combines both human holistic scores of essays and approximations to human holistic score generated by e-rater® version 2.0 (Attali, Burstein, & Andreyev, 2003; Burstein, Chodorow, & Leacock, in press). It is also possible to report essay features on their own or to develop analytic essay scores based on some combination of essay features. For this purpose, it is important to examine the statistical and measurement properties of essay features that are employed directly or indirectly in essay assessment. Essay features considered are described in Section 1. In that section, statistical properties considered include means, standard deviations, coefficients of skewness, and coefficients of kurtosis. These measures are employed to suggest nonlinear transformations that appear meaningful and that have distributions that are more nearly normal. Although analysis in this report is not primarily intended to consider prediction of human holistic scores by computer-generated essay features, some consideration will be given to correlations between computer-generated essay features and human holistic scores. Elementary statistical models are used to suggest inherent limitations in many of the numerical features under study. Measurement properties studied involve reliability of raters in the case of holistic scoring and variability of essay features by examinee and by prompt.

Analysis of data is based on two prompts from the GMAT exam, two prompts from the TOEFL exam, and a larger number of prompts used to generate essays via Criterion^{SM}. To avoid analysis of texts that cannot reasonably be regarded as essays, submitted texts of fewer than 25 words are discarded.

The essays from the GMAT and TOEFL examinations provided data on rater agreement and on variability of distributions of numerical features by program and prompt (Section 2).

The essays from Criterion provide further data on variability of distributions by population and prompt. They include multiple essays by the same individual, so that they provide information concerning reliability of features in assessment of individual examinees (Section 3). These data concerning reliability also provide some information concerning limits on the reliability of scales constructed from linear combinations of features (Section 4).

Implications of this report are considered in Section 5.

1 Common Features and Their Distributions

The features considered in this report include human holistic scores, features used in e-rater version 2.0 (Attali et al., 2003; Burstein et al., in press), features generated by the style program included in the diction-1.02 package distributed by the GNU project, and several features computed by the author by use of perl and Fortran 95 programs. When essentially the same feature was available in both e-rater version 2.0 and another source, e-rater was employed in analysis.

Human Holistic Scores

Human holistic scores are obtained from trained readers in a variety of common examinations. Examples include essays in GMAT, TOEFL, GRE, and SAT II Writing. In these examinations, essays are read and assigned an integer score from 1 to 6 based on a rubric used to describe the meaning of the scores. Different raters who look at the same essay do not necessarily assign the same rating, so that more than one rater is often employed, and some procedure is employed to supply a summary holistic score based on the individual ratings. For each of the GMAT and TOEFL prompts under study in this report, for any essay written by an examinee, two raters provide a holistic score for the essay. A third holistic score is supplied by a third rater if the first two holistic scores differ by two or more. For elementary analysis, issues of selection bias and missing data are most easily avoided by examination of the basic statistical properties of the holistic scores supplied for each essay by the initial two raters. Such properties are summarized in Tables 1 and 2 for each GMAT and TOEFL prompt under study.

Table 1.
Summary Statistics for Average Holistic Score

Program	Prompt	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	5183	3.789	1.118	-0.143	-0.450
GMAT	2	5158	3.842	1.067	-0.119	-0.319
TOEFL	1	4895	4.112	1.039	-0.161	-0.295
TOEFL	2	4884	4.075	1.021	-0.112	-0.238

Table 2.
Summary Statistics for Difference of Holistic Scores

Program	Prompt	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	5183	0	0.844	0	0.515
GMAT	2	5158	0	0.832	0	0.122
TOEFL	1	4895	0	0.742	0	0.292
TOEFL	2	4884	0	0.719	0	0.438

In Table 1, summary statistics are computed by use of standard unbiased estimates of central moments (Cramér, 1946). In Table 2, the summary statistics are computed based on the assumption that the distribution of the first two human scores must be symmetric, for the label of first versus second score is arbitrary. The symmetry assumption implies that the mean and skewness must be 0, the estimated standard deviation is the square root of the average of the squared differences between the two scores, and the kurtosis is computed with the fourth central moment estimated by the average of the fourth powers of the score differences.

The results suggest that the distributions of the average holistic score and of the difference of holistic scores have no remarkable properties. The slight negative skewness is predictable given that the mean is closer to the upper bound of the scores than to the lower bound of the scores. The variability of the raters is quite substantial. The estimated intraclass correlations for the four prompts are 0.750, 0.736, 0.774, and 0.779, respectively.

Testing programs vary in terms of methods used to summarize holistic scores. SAT II Writing in effect uses average holistic score. If the initial scores do not differ by more than

1, then the average holistic score is used by GRE and TOEFL; however, these programs employ at least one additional reader if the initial scores differ by more than 1. Prior to adoption of e-rater by GMAT, a similar approach was used by that program. In TOEFL, when the initial scores differ by more than one, a table leader grades the essay and assigns the score, so that initial scores are not used. This approach is not confined to ETS. At ACT, it is used in the Collegiate Assessment of Academic Proficiency (CAAP) Writing Essay. One approach, Method 1, is to report the average holistic score if the third reader provides a score exactly equal to this average and otherwise to report the average of the score from the third reader and from the initial reader with a score closest to that of the third reader. This approach is used by the GRE unless the third reader has a score that differs from each previous reader by at least 2. In that case, a fourth reader is used. If the fourth reader has a score that differs by one from two different previous scores, then the score from the fourth reader is used. Otherwise, the final score is the average of the score of the fourth reader and the score of the reader that only differs from the fourth reader by one. (If a fourth reader is used, then the fourth reader cannot have a score that differs by two or more from that of any of the three previous readers.) A variant on Method 1 was used with the College Board Admissions Testing Program's English Composition Achievement Test with Essay (Mazzeo, Schmit, & Cook, 1986a; Mazzeo, Schmit, & Cook, 1986b). The change was that a third reader was only used if the score difference between the first two raters was at least 3. The rather straightforward approach, Method 2, in which the three scores are averaged, is not used by any testing program known to the author, although a variant of this approach has been studied in which a third reader is only used if initial scores differ by at least three and the average of the three scores is then rounded to the nearest half if the initial scores differ by at least three (Mazzeo et al., 1986a; Mazzeo et al., 1986b).

For the prompts under study, Method 1 results in slightly larger standard deviations than does use of the average holistic score, but the differences are modest, for standard deviation is never increased by more than 2% for any of the four prompts. Method 2 results in standard deviations quite similar to those for average holistic score. The percentage change ranges from -0.001 to 0.003. There does not seem to be a compelling statistical argument against use of average holistic score without ever using a third rater. This result

is consistent with previous work on the Admissions Testing Program's English Composition Test with Essay (Mazzeo et al., 1986a; Mazzeo et al., 1986b). Thus the SAT practice is readily justified. On the other hand, a decision to use Method 1 or Method 2 does not appear to cause significant harm, so that a decision to use these methods or methods used by TOEFL or GRE to improve public perception of the scoring system can be justified. More restricted use of third readers is also an option.

For the prompts from GMAT and TOEFL under study, a resolved holistic score is used by e-rater. This resolved score is always an integer between 1 and 6, so the approach used by e-rater necessarily differ from any of the previously considered scoring methods. The resolved score is computed by the following procedure. If the initial two ratings do not differ by more than 1, then the resolved score is the maximum of the two ratings. If the two ratings differ by more than 1, then the resolved holistic score is the holistic score provided by a third rater. The fraction of essays with resolved holistic score equal to the maximum of the two initial holistic scores is quite high. For the GMAT prompts, the fractions are 0.964 and 0.975. For the TOEFL prompts, the fractions are 0.987 and 0.988. The difference between resolved holistic score and the maximum of the two initial scores has means for the four prompts that range from -0.011 to 0.005 and standard deviations that range from 0.116 to 0.207. As evident from Table 3, the resolved holistic score has a higher standard deviation and a higher mean than the average holistic score. These results are hardly surprising, for a maximum normally has a higher mean and a higher standard deviation than an average. In practice, resolved holistic score and average holistic score are highly correlated. For the prompts considered, all estimated correlation coefficients are between 0.966 and 0.967. Thus resolved holistic score appears somewhat less desirable than is average holistic score.

Essay Length

The length of an essay is an elementary measure of the essay's development, especially to the extent that a very short essay cannot provide a sufficiently detailed response to a prompt. The simplest measures of length are the number of words in an essay and the number of characters in an essay. The number of words in an essay is a readily computed

Table 3.
Summary Statistics for Resolved Holistic Score

Program	Prompt	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	5183	4.004	1.173	-0.179	-0.497
GMAT	2	5158	4.074	1.106	-0.097	-0.368
TOEFL	1	4895	4.322	1.067	-0.208	-0.334
TOEFL	2	4884	4.275	1.058	-0.156	-0.300

measure of essay length that has been studied for some time as a predictor of holistic score (Page, 1967; Reid & Findlay, 1986). This number is computed by both e-rater and style, and e-rater applies this variable in its regression analysis. The command "wc" in UNIX also provides a word count. Slight discrepancies between computations reflect decisions concerning handling of hyphens, apostrophes, and numerals. For instance, e-rater and style both regard hyphens and apostrophes as separating words, so that "Bob's" and "part-time" are both rendered as two words. The UNIX command wc regards "Bob's" and "part-time" as single words. With all programs, typographical errors may change word counts when spaces are omitted. Thus typing "wenthome" rather than "went home" reduces the word count by one. In addition, for wc, an error in which spacing is omitted after a punctuation mark can also reduce the word count.

A more significant problem is that the style program functions somewhat better with professionally written text than with texts produced by students with an imperfect command of the English language. The style program does not count words in passages that are not considered portions of sentences, and the rules for identification of sentences require some conformance with customary standards for capitalization at the beginning of sentences and punctuation at the completion of sentences. Indeed, sufficiently unorthodox capitalization and punctuation can result in no program output at all. This issue is not entirely theoretical, especially for the TOEFL examination. For the first prompt from this examination, 34 essays were not processed. For the other prompt, 18 essays were not processed. Even for the GMAT essays, 15 essays were unprocessed for the first prompt, and 9 for the second prompt.

Despite all these issues, the word counts from e-rater and from style agree exactly for each prompt at least in 65% of all essays, with agreement rates of 74% and 80% in the two GMAT essays. In terms of sample correlations of the two measures of word count for essays processed by style, correlations for GMAT prompts were 0.991 and 0.994, while correlations as TOEFL prompts were 0.974 and 0.977. Results for word counts for wc and e-rater are very close, for sample correlations for GMAT prompts are at least 0.999 and sample correlations for TOEFL prompts are at least 0.999. Nonetheless, the problem with processing of unconventionally written sentences and the problem with unconventional spacing made use of the word count from e-rater the appropriate choice in this report.

It is common statistical practice to consider nonlinear transformations of variables (Scheffé, 1959, section 10.7). Such transformations are generally considered in order to yield a variable with a distribution more similar to a normal distribution, to yield a variable with a variance that is relatively unrelated to its mean, or to yield a regression with another variable that is more nearly linear. For a positive variable such as the word count of an essay, the power family of transformations is commonly considered. Simple relevant cases include the square root and logarithmic transformations. In addition to the previously described statistical considerations, use of square roots or logarithms in an assessment scale would have the effect that addition of a few words to a short essay would have more impact on the quality of the essay than would the addition of a few words to a long essay. This issue influences the use of the fourth root of word count in Project Essay Grader (Page, 1994). Logarithms have some further attraction in terms of other variables under study such as number of characters and average word length in that they facilitate many decompositions of interest. With these considerations in mind, consider the results of Table 4. The fourth root was also examined, but results are not presented because no clear gain was found from this transformation relative to square roots and logarithms.

In terms of distributions, the square root transformation has the advantage of small skewness and relatively modest kurtosis, so that the distribution of square roots of word counts is relatively close to a normal distribution. The logarithm does not have an advantage in terms of distribution compared to the original scale. On the whole, the original scale appears preferable to the logarithmic one in terms of skewness and kurtosis.

Table 4.
Summary Statistics for Transformations of Number of Words

Program	Prompt	Function	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	Identity	5183	291.22	101.513	0.433	0.357
GMAT	1	Root	5183	16.793	3.038	-0.158	0.249
GMAT	1	Log	5183	5.606	0.388	-0.923	2.045
GMAT	2	Identity	5158	304.744	107.326	0.624	0.733
GMAT	2	Root	5183	17.181	3.089	0.019	0.303
GMAT	2	Log	5183	5.653	0.378	-0.733	1.709
TOEFL	1	Identity	4895	232.906	84.380	0.747	1.425
TOEFL	1	Root	4895	15.008	2.766	0.061	0.497
TOEFL	1	Log	4895	5.381	0.388	-0.733	1.709
TOEFL	2	Identity	4884	244.657	88.992	0.785	1.201
TOEFL	2	Root	4884	15.384	2.829	0.132	0.459
TOEFL	2	Log	4884	5.431	0.384	-0.678	1.846

For each transformation, the correlations with average holistic score are quite high, especially if the correction for attenuation is employed to examine the anticipated result of use of an arbitrarily large number of randomly chosen raters rather than just two (Gulliksen, 1987, pp. 101–104). Results are summarized in Table 5.

In terms of correlations, square roots and logarithms of the number of words are a bit more highly correlated with average holistic score than is number of words itself. With the adjustment for attenuation, the correlations are strikingly high.

The number of alphanumeric characters in an essay provides an alternative length measure that is easily computed. This variable is computed by both e-rater and style. The character count in wc includes punctuation, spaces, and line returns, so that it is less satisfactory. As in the case of number of words, there are modest variations present in results from the two programs, with significant differences encountered when sentences do not have proper capitalization and punctuation. Once again, use of e-rater computations appears preferable. The e-rater version 2.0 regression does not use number of characters directly, although this variable is employed in intermediate computations. Transformation considerations for number of characters are essentially similar to those for number of words. As evident from Tables 6 and 7, results for number of characters are quite similar

Table 5.
Correlations of Average Holistic Score and Transformations of Number of Words

Test	Prompt	Function	Count	Sample correlation	Corrected correlation
GMAT	1	Identity	5183	0.790	0.854
GMAT	1	Root	5183	0.811	0.864
GMAT	1	Log	5183	0.791	0.854
GMAT	2	Identity	5158	0.811	0.880
GMAT	2	Root	5158	0.824	0.894
GMAT	2	Log	5158	0.817	0.887
TOEFL	1	Identity	4895	0.774	0.828
TOEFL	1	Root	4895	0.801	0.858
TOEFL	1	Log	4895	0.808	0.865
TOEFL	2	Identity	4884	0.768	0.820
TOEFL	2	Root	4884	0.793	0.848
TOEFL	2	Log	4884	0.800	0.855

to those for number of words, although transformations based on the number of characters are even more highly correlated with average holistic score than are the corresponding transformations based on number of words. Results are particularly extreme for the square root transformation for the second GMAT prompt, for the correction for attenuation leads to an estimated correlation of 0.932 between the square root of the number of characters and the average essay score for an arbitrarily large number of randomly selected raters.

Obviously transformations based on the number of words are very highly correlated with transformations based on the number of characters, especially if the transformations are comparable. For the prompts studied, the sample correlations for all pairs in which one transformation is based on number of words and one is based on the number of characters is at least 0.938. For cases in which the same transformation is used for both word and character counts, the sample correlation is at least 0.981.

Given the results concerning the strong relationship of number of characters with holistic score, subsequent variables will be examined in light of the relationship. In tabular presentations, the conventions will be adopted that average holistic score is denotes by AHS and square root of number of characters by SCH.

Table 6.
Summary Statistics for Transformations of Number of Characters

Program	Prompt	Function	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	Identity	5183	1439.840	501.035	0.400	0.292
GMAT	1	Root	5183	37.332	6.762	-0.184	0.228
GMAT	1	Log	5183	6.7626	0.389	-0.938	2.004
GMAT	2	Identity	5158	1463.120	510.790	0.561	0.573
GMAT	2	Root	5183	37.652	6.741	-0.031	0.269
GMAT	2	Log	5183	7.223	0.378	-0.781	1.788
TOEFL	1	Identity	4895	1108.310	400.180	0.690	1.224
TOEFL	1	Root	4895	32.739	6.039	0.018	0.448
TOEFL	1	Log	4895	6.941	0.390	-0.776	1.797
TOEFL	2	Identity	4884	1057.840	387.050	0.776	1.183
TOEFL	2	Root	4884	31.980	5.926	0.118	0.453
TOEFL	2	Log	4884	6.894	0.388	-0.702	1.902

Table 7.
Correlations of Average Holistic Score and Transformations of Number of Characters

Test	Prompt	Function	Count	Sample correlation	Corrected correlation
GMAT	1	Identity	5183	0.808	0.873
GMAT	1	Root	5183	0.816	0.882
GMAT	1	Log	5183	0.806	0.870
GMAT	2	Identity	5158	0.847	0.920
GMAT	2	Root	5158	0.858	0.932
GMAT	2	Log	5158	0.849	0.922
TOEFL	1	Identity	4895	0.812	0.869
TOEFL	1	Root	4895	0.836	0.894
TOEFL	1	Log	4895	0.838	0.897
TOEFL	2	Identity	4884	0.804	0.859
TOEFL	2	Root	4884	0.828	0.885
TOEFL	2	Log	4884	0.831	0.888

Table 8.
Summary Statistics for Average Word Length

Program	Prompt	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	5183	4.947	0.237	-0.141	0.253
GMAT	2	5158	4.814	0.326	0.123	0.142
TOEFL	1	4895	4.767	0.305	0.077	0.067
TOEFL	2	4884	4.328	0.281	0.175	0.055

Distribution of Word Length.

The distribution of word length has been studied as a measure of sophistication of the vocabulary used in an essay. Measures of this distribution that have been considered in the literature have been average word length and standard deviation of word length.

Average word length. Average word length is especially easy to consider, for it is the ratio of the number of characters to the number of words. Average word length has been employed in the Coleman-Liau formula for readability (Coleman & Liau, 1975) and in the automated readability index (Smith & Kincaid, 1970). The relationship of average word length to holistic scoring has been studied for some time (Reid & Findlay, 1986). Both e-rater and style compute this variable, and e-rater version 2.0 applies average word length in its regression analysis. Variations between e-rater and style in their measurement of this variable are minimal. Values reported here are those for e-rater. Average word length presents no difficulties in terms of distribution, as is evident from Table 8. As a consequence, there is no obvious gain from use of transformations.

For the prompts in Table 8, average word length has a modest correlation with human score and a rather low correlation with the logarithm or square root of the number of characters. To illustrate results, consider Tables 9 and 10.

Corrections for attenuation for the correlation of average holistic score and average word length lead to the respective values 0.144, 0.185, 0.217, and 0.249 for the four prompts. Thus average word length is not a major predictor of holistic scoring.

Some variations on average word length can be considered in which selected words are removed from the computation. One approach is to remove stop words such as "the"

Table 9.
Correlations for Average Word Length

Program	Prompt	Count	Correlation AHS	Correlation SCH	Corrected with AHS
GMAT	1	5183	0.134	0.081	0.144
GMAT	2	5158	0.170	0.077	0.185
TOEFL	1	4895	0.202	0.103	0.217
TOEFL	2	4884	0.233	0.135	0.249

Table 10.
Partial Correlation of Average Word Length and Average Holistic Score Given Square Root of Number of Characters

Program	Prompt	Count	Partial correlation Sample	Partial correlation Corrected
GMAT	1	5183	0.117	0.155
GMAT	2	5158	0.204	0.313
TOEFL	1	4895	0.213	0.280
TOEFL	2	4884	0.218	0.280

or "of" from calculations. Another approach is to remove both stop words and words in the prompt. The list of stop words to exclude was the same as the ones used in e-rater. Other lists are available. These variations on average word length do not appear to be available directly from either the e-rater or style software, so they were computed by the author by use of a standard perl program for compilation of word frequencies of a text (http://dada.perl.it/shootout/wordfreq.perl.html) and a Fortran 95 program to compute lengths of words and to compute average word length, among other statistics. These approaches emphasize word lengths for more significant words. The variations on average word length do not cause any significant problems in terms of distributions. For the four prompts and the two approaches, no estimated coefficient of skewness or kurtosis exceeds 1 in magnitude, and the majority of estimates are rather close to 0. On the other hand, partial correlations of average human score and modified average word length given square root of number of characters remain relatively small, although the partial correlations exceed those for average word length with all words included. With stop words removed, the estimated partial correlations not corrected for attenuation range from 0.151 to 0.278

Table 11.
Summary Statistics for Standard Deviation of Word Length

Program	Prompt	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	5183	2.619	0.150	-0.091	0.549
GMAT	2	5158	2.635	0.223	-0.070	0.317
TOEFL	1	4895	2.620	0.220	-0.132	0.112
TOEFL	2	4884	2.248	0.231	0.047	0.036

Table 12.
Correlations for Standard Deviation of Word Length

Program	Prompt	Count	Correlation AHS	Correlation SCH	Corrected with AHS
GMAT	1	5183	0.092	0.085	0.099
GMAT	2	5158	0.274	0.160	0.298
TOEFL	1	4895	0.232	0.114	0.249
TOEFL	2	4884	0.233	0.285	0.305

for the four prompts. With stop words and prompt words removed, the corresponding range is from 0.173 to 0.272.

Standard deviation of word length. The standard deviation of word length has been considered as a predictor of human holistic score (Page, 1967). The rationale for this measure is somewhat similar to that for average word length. As evident from the data in Tables 11, 12, and 13, the standard deviation of word length is quite competitive with average word length. Nonetheless, the relationship with holistic score is still a modest one.

Table 13.
Partial Correlation of Standard Deviation of Word Length and Average Holistic Score Given Square Root of Number of Characters

Program	Prompt	Count	Partial correlation Sample	Partial correlation Corrected
GMAT	1	5183	0.039	0.052
GMAT	2	5158	0.271	0.416
TOEFL	1	4895	0.251	0.330
TOEFL	2	4884	0.268	0.344

Average syllables per word. A related measure to average word length is average number of syllables per word. This measure is computed in the style program, and it is employed in the Flesch and Kincaid readability indexes (Flesch, 1948; Kincaid, Fishburne, Rogers, & Chissom, 1975). The style program seeks to ascertain the number of syllables per word by an examination of consonants and vowels in the word rather than by use of a dictionary, so that some errors are to be expected. For instance, both "Mary" and "little" are regarded as words with one syllable. For prompts processed by style, average number of syllables had a somewhat higher kurtosis than average word length, and correlations of average number of syllables with average holistic score were comparable to those for average word length and average holistic score. Given that average word length is easier to compute and more accurately computed than is average number of syllables per word, there appears little reason to prefer average number of syllables per word to average word length.

Readability measures. Given the role of average word length and average number of syllables per word in readability formulas, one might reasonably ask whether readability indexes have any advantage over average word length. Readability indexes have been obtained by a number of researchers in an effort to use regression analysis to predict the difficulty of a passage. These measures may reflect textual complexity. In principle, there might be some gain because the readability indexes also consider a measure of average sentence length; however, in fact, readability indexes are quite disappointing. The Coleman-Liau formula for readability (Coleman & Liau, 1975) computed by style is $5.89W - 30/S - 15.8$, where W is average word length and S is average number of words per sentence. For each of the four prompts, the Coleman-Liau readability measure computed by style is so highly correlated with average word length that little gain is possible. Even for the e-rater computations of average word length, correlations are at least 0.98 for each prompt.

The automated readability index (Smith & Kincaid, 1970) computed by the style program is $4.71W + 0.5S - 21.43$. This index is remarkably unsatisfactory in the context of essay assessment, evidently due to its relatively strong dependence on sentence length. For each prompt, the index has a very high kurtosis, a weak correlation with average word length, and very little relationship with average holistic score. An underlying issue is that,

among essays processed by style, the relationships of average sentence length to average word length and to average holistic score are both weak, and the distribution of average sentence length has a very large skewness and a very large kurtosis for each prompt. It also follows that other readability measures that depend strongly on average sentence length are rather unsatisfactory. This issue affects the Flesch index (Flesch, 1948) and the Flesch-Kincaid index (Kincaid et al., 1975) computed by style. If Y is the average number of syllables per word, then the Flesch index is $206.835 - 84.6Y - 1.015S$ and the Flesch-Kincaid index is $11.8Y + 0.39S - 15.59$.

In summary, competitors do exist for the average word length currently in the regression model in e-rater version 2.0. Nonetheless, none of the measures considered related to distribution of word length is a major predictor of holistic score.

Grammar, Usage, Mechanics, and Style

In essay assessment, measures may be considered for attributes such as grammar, usage, mechanics, and style. Such measurement is used in the regression analysis in e-rater version 2.0 (Attali et al., 2003; Burstein et al., in press). The software provides a count of 33 separate types of errors, of which nine involve grammar, seven involve usage, 11 involve mechanics, and six involve style. In the regression analysis in e-rater version 2.0, the actual variables used are the total number of grammar errors divided by the number of words, the total number of usage errors divided by the number of words, the total number of mechanics errors divided by the number of words, and the total number of style errors divided by the number of words. The division by number of words reflects the increasing opportunity to err presented by essays of increasing length.

In general, no comparable analysis of these aspects of essays appears to be available from other software, especially in the public domain. The most important exception involves spelling errors, which are classified as errors in mechanics in e-rater. Spell checking in e-rater relies on the program spell available from the GNU project. This utility is in the process of being replaced in GNU by aspell and pspell. A similar change is to take place in later versions of e-rater. Spelling has been considered as a predictor of human holistic score for some time (Page, 1968; Reid & Findlay, 1986). The program style treats short sentences,

Table 14.
Relative Frequencies of Recorded Grammar, Usage, Mechanics, and Style Errors

Program	Prompt	Count	GU	Spelling	OM	REP	OS
GMAT	1	5183	0.047	0.307	0.029	0.598	0.019
GMAT	2	5158	0.067	0.222	0.034	0.662	0.016
TOEFL	1	4895	0.074	0.229	0.049	0.629	0.020
TOEFL	2	4884	0.056	0.136	0.041	0.743	0.024

long sentences, and passive sentences, three of the error criteria considered in e-rater, as style errors. The program diction in the GNU project provides some further information concerning errors in style, although this information is relatively limited. Some variables from an earlier version of diction have been considered as predictors of holistic scores (Reid & Findlay, 1986). The proprietary program IntelliMetric™ purports to employ features that are described as mechanical in its publications, but no detail is available (Elliot, 2003). A similar issue applies to the Intelligent Essay Assessor™ (Landauer, Laham,& Foltz, 2003).

Although the summary variables used in the e-rater regression are reasonable, it is still appropriate to consider alternative summaries. Two issues arise. One involves the very unequal frequencies of different kinds of errors. The second involves the usual questions concerning distributions in the original scale and in transformed scales.

The variation in error frequency is rather striking. A remarkable fraction of the observed errors involve spelling errors and words tagged as repetitive. Consider Table 14. In this and later tables, GU is grammar and usage, OM is mechanics other than spelling, REP is repetitive words, and OS is style other than repetitive word.

As evident from Table 15, errors other than spelling and repetitive word usage are not especially common. The average number of other errors per essay ranges from 2.51 for the first GMAT prompt to 4.10 for the second TOEFL prompt. In addition to the information provided by Table 15, it is notable that the fraction of essays without any reported style error other than repetitive words ranges from 0.670 for the first GMAT prompt to 0.780 for the second GMAT prompt. In the case of errors in mechanics other than spelling, the

Table 15.
Average Frequencies per Essay of Recorded Grammar, Usage, Mechanics, and Style Errors

Program	Prompt	Count	GU	Spelling	OM	REP	OS
GMAT	1	5183	1.245	8.128	0.770	15.844	0.497
GMAT	2	5158	1.570	5.239	0.795	15.621	0.367
TOEFL	1	4895	1.905	5.919	1.271	16.294	0.508
TOEFL	2	4884	1.895	4.601	1.379	25.199	0.829

fraction of essays with no errors ranges from 0.418 for the second TOEFL prompt to 0.579 for the first GMAT prompt. For all cases of errors in grammar or usage, the fractions of essays with no errors ranges from 0.260 for the first TOEFL prompt to 0.381 for the first GMAT prompt.

Distributions of error rates create additional problems. Because rates cannot be negative and can be much larger than the average in some instances, it is to be expected that skewness and kurtosis will be relatively large. This expectation is indeed met. Consider Table 16.

Tables 14 and 16 suggest considerable difficulties in application of the grammar, usage, mechanics, and style information. Variables that are 0 for substantial fractions of essays are not likely to have much discriminatory power, and very large coefficients of skewness and kurtosis suggest major issues with outliers.

As evident from Table 17, the most serious problems with distributions may be solved by use of square root transformations. These transformations do not remove the problem that several variables are often 0.

Grammar, usage, mechanics, and style variables do have some relationship to holistic scores; however, considerable care is needed to interpret results. Consider Tables 18 and 19.

There is a striking difference between partial correlations and ordinary correlations in the cases of repetitive words and total errors. A possible issue is that the definition of repetitive words used in e-rater depends indirectly on the total number of words in the essay (Burstein et al., in press). Because such a large fraction of errors in essays are repetitive words, the results for total errors are also affected. A further unusual result involves the

Table 16.
Distributions of Error Rates

Program	Prompt	Rate	Count	Mean	Standard deviation	Skewness	Kurtosis
GMAT	1	Total	5183	0.102	0.066	1.352	3.327
GMAT	1	GU	5183	0.005	0.006	2.776	16.355
GMAT	1	Spelling	5183	0.029	0.020	1.558	5.162
GMAT	1	OM	5183	0.003	0.006	5.099	37.628
GMAT	1	REP	5183	0.063	0.055	1.256	2.379
GMAT	1	OS	5183	0.002	0.004	7.747	130.162
GMAT	2	Total	5158	0.090	0.069	1.414	2.791
GMAT	2	GU	5158	0.006	0.006	2.997	18.973
GMAT	2	Spelling	5158	0.019	0.017	2.612	13.845
GMAT	2	OM	5158	0.003	0.006	4.503	31.831
GMAT	2	REP	5158	0.062	0.061	1.551	3.607
GMAT	2	OS	5158	0.001	0.003	5.729	47.534
TOEFL	1	Total	4895	0.126	0.084	1.330	2.845
TOEFL	1	GU	4895	0.009	0.009	1.954	6.888
TOEFL	1	Spelling	4895	0.028	0.025	2.048	7.253
TOEFL	1	OM	4895	0.006	0.010	4.682	38.233
TOEFL	1	REP	4895	0.081	0.069	1.393	3.168
TOEFL	1	OS	4895	0.002	0.007	5.422	37.334
TOEFL	2	Total	4884	0.155	0.087	1.136	2.404
TOEFL	2	GU	4884	0.008	0.009	2.055	7.026
TOEFL	2	Spelling	4884	0.021	0.020	2.974	18.818
TOEFL	2	OM	4884	0.006	0.012	5.577	51.109
TOEFL	2	REP	4884	0.116	0.076	1.010	1.584
TOEFL	2	OS	4884	0.003	0.010	4.433	25.815

Table 17.
Distributions of Square Roots of Error Rates

Program	Prompt	Rate	Count	Mean	Standard deviation	Skewness	Kurtosis
GMAT	1	Total	5183	0.303	0.101	0.303	0.162
GMAT	1	GU	5183	0.051	0.046	0.419	-0.419
GMAT	1	Spelling	5183	0.160	0.059	0.034	0.823
GMAT	1	OM	5183	0.033	0.044	1.370	2.287
GMAT	1	REP	5183	0.216	0.130	-0.267	-0.478
GMAT	1	OS	5183	0.023	0.035	1.493	2.713
GMAT	2	Total	5158	0.279	0.112	0.329	-0.020
GMAT	2	GU	5158	0.058	0.047	0.313	-0.200
GMAT	2	Spelling	5158	0.122	0.060	0.374	1.100
GMAT	2	OM	5158	0.033	0.043	1.189	1.338
GMAT	2	REP	5158	0.206	0.140	-0.040	-0.625
GMAT	2	OS	5158	0.015	0.031	2.210	5.242
TOEFL	1	Total	4895	0.336	0.116	0.247	0.205
TOEFL	1	GU	4895	0.076	0.055	0.087	-0.521
TOEFL	1	Spelling	4895	0.151	0.072	0.313	0.596
TOEFL	1	OM	4895	0.054	0.057	0.953	1.248
TOEFL	1	REP	4895	0.250	0.137	-0.220	-0.141
TOEFL	1	OS	4895	0.020	0.042	2.444	6.559
TOEFL	2	Total	4884	0.377	0.111	0.143	0.314
TOEFL	2	GU	4884	0.074	0.054	0.149	-0.443
TOEFL	2	Spelling	4884	0.066	0.362	2.974	1.227
TOEFL	2	OM	4884	0.055	0.058	1.166	2.485
TOEFL	2	REP	4884	0.318	0.122	-0.390	0.686
TOEFL	2	OS	4884	0.026	0.052	2.201	4.636

Table 18.

Correlations for Square Roots of Grammar, Usage, Mechanics, and Style Variables With Average Holistic Score and the Square Root of the Number of Characters

Program	Prompt	Variable	Count	Correlation		Corrected with AHS
				AHS	SCH	
GMAT	1	Total	5183	-0.456	-0.472	-0.493
GMAT	1	GU	5183	-0.201	-0.142	-0.218
GMAT	1	Spelling	5183	-0.124	-0.130	-0.134
GMAT	1	OM	5183	-0.168	-0.125	-0.181
GMAT	1	REP	5183	-0.404	-0.430	-0.436
GMAT	1	OS	5183	0.010	0.046	0.011
GMAT	2	Total	5158	-0.589	-0.564	-0.640
GMAT	2	GU	5158	-0.201	-0.114	-0.218
GMAT	2	Spelling	5158	-0.306	-0.206	-0.332
GMAT	2	OM	5158	0.170	-0.145	0.185
GMAT	2	REP	5158	-0.493	-0.512	-0.536
GMAT	2	OS	5158	0.019	0.057	0.020
TOEFL	1	Total	4895	-0.623	-0.542	-0.667
TOEFL	1	GU	4895	-0.309	-0.152	-0.330
TOEFL	1	Spelling	4895	-0.413	-0.274	-0.442
TOEFL	1	OM	4895	-0.259	-0.161	-0.277
TOEFL	1	REP	4895	-0.473	-0.467	-0.506
TOEFL	1	OS	4895	-0.013	0.049	-0.013
TOEFL	2	Total	4884	-0.603	-0.568	-0.644
TOEFL	2	GU	4884	-0.275	-0.123	-0.293
TOEFL	2	Spelling	4884	-0.413	-0.205	-0.442
TOEFL	2	OM	4884	-0.250	-0.157	-0.268
TOEFL	2	REP	4884	-0.486	-0.519	-0.519
TOEFL	2	OS	4884	-0.020	0.046	-0.021

Table 19.
Partial Correlations of Square Roots of Grammar, Usage, Mechanics, and Style Variables With Average Human Score Given Square Root of the Number of Characters

				Partial correlation	
Program	Prompt	Variable	Count	Sample	Corrected
GMAT	1	Total	5183	-0.139	-0.184
GMAT	1	GU	5183	-0.201	-0.309
GMAT	1	Spelling	5183	-0.032	-0.042
GMAT	1	OM	5183	-0.115	-0.152
GMAT	1	REP	5183	-0.101	-0.134
GMAT	1	OS	5183	-0.047	-0.062
GMAT	2	Total	5158	-0.248	-0.381
GMAT	2	GU	5158	-0.201	-0.309
GMAT	2	Spelling	5158	-0.256	-0.394
GMAT	2	OM	5158	-0.200	-0.307
GMAT	2	REP	5158	-0.123	-0.189
GMAT	2	OS	5158	-0.059	-0.091
TOEFL	1	Total	4895	-0.369	-0.485
TOEFL	1	GU	4895	-0.334	-0.440
TOEFL	1	Spelling	4895	-0.348	-0.457
TOEFL	1	OM	4895	-0.230	-0.302
TOEFL	1	REP	4895	-0.170	-0.224
TOEFL	1	OS	4895	-0.114	-0.150
TOEFL	2	Total	4884	-0.287	-0.369
TOEFL	2	GU	4884	-0.310	-0.398
TOEFL	2	Spelling	4884	-0.303	-0.389
TOEFL	2	OM	4884	-0.217	-0.279
TOEFL	2	REP	4884	-0.118	-0.151
TOEFL	2	OS	5183	-0.102	-0.132

first GMAT prompt. The spelling variable has an ordinary and partial correlation that is unusually small in magnitude. This problem may reflect difficulties with the spell checker used. The first essay has a prompt that uses the place name "Cambria." This name is not recognized by the spell checker as a valid word. As a consequence, many essays in which writers have engaged in the perfectly reasonable practice of mentioning the place in their response have been penalized. Indeed, the simple expedience of removing from the list of spelling errors all uses of the word "Cambria" changes the correlation of average holistic score and square root of spelling error rate to -0.476. This issue is particularly worthy of note because it is quite easy to add a supplementary dictionary to a spell checker to include any prompt word not normally accepted. With the current dictionary, the only word in the four prompts that is affected is Cambria. The word "Cambria" remains a problem even with the newer aspell utility unless the largest available dictionary is used.

The problem observed with the spelling variable raises basic questions concerning all grammar, usage, mechanics, and style variables. To what extent do these variables reflect peculiarities of computer programs? Were very careful human scoring of these variables used, what changes would occur in the values of the variables and in their relationships to other relevant variables such as essay length and human holistic score? Do the differences between computer-generated variables and human-generated variables have any tendency to affect different classes of examinees in different fashions? To what extent are errors in grammar, usage, mechanics, and style the result of typographical errors overlooked by the examinee? To the extent that measurement errors are substantial, it is reasonable to expect that correlations of measures of grammar, usage, mechanics, and style with other variables will be decreased to a substantial extent.

Accuracy of the identification of errors in e-rater has indeed been studied (Burstein et al., in press); however, variations in accuracy by prompt, subject, or program have not been considered to date. Thus the questions raised have not yet been addressed directly. Nonetheless, it is clear that error rates are not negligible. No reason exists to believe that this situation is unique to the software used in e-rater (Leacock & Chodorow, 2003).

The existence of deviations between computer analysis and expert human analysis may affect public perceptions of automatic scoring, although it is worth emphasizing that

moderate error rates by the software need not have major adverse impact on the use of the grammar, usage, mechanics, and style measures, provided that the manner in which software and human experts differ does not depend on the essay features of main interest. This point may be made by a rather oversimplified mathematical model in which, for a randomly chosen essay, N is the essay length in words, Z is an essay feature of interest, say average holistic score, P is an error rate for a particular kind of error that occurs for individual words in an essay, and, given N, P, and Z, (X_i, Y_i), $1 \leq i \leq N$, are independent and identically distributed pairs of random variables, each of which has value 0 or 1. The variable X_i is 1 only if the kind of writing error under study exists for the ith word in the essay, and Y_i is 1 only if the computer program identifies that error for word i. Given N, P, and Z, the conditional probability that $X_i = 1$ is P. It is assumed that identification of errors is unrelated to N, P, and Z in the sense that, given $X_i = k$, the conditional probability that $Y_i = 1$ is p_k for a constant p_k greater than 0 and less than 1. The constant p_k does not depend on the covariate Z, the word count N, or the error rate P. The true error rate for the essay is $T = N^{-1} \sum_{i=1}^{N} X_i$, and the observed rate is $O = N^{-1} \sum_{i=1}^{N} Y_i$. For simplicity, the possibility of not counting words correctly is ignored. The conditional expected value of T given N, Z, and P is P, so that the covariance of Z and T is the same as the covariance C_{ZP} of Z and P. The conditional variance of T is $P(1-P)/N$, so that the unconditional variance $\sigma^2(T)$ of T is the sum of the variance $\sigma^2(P)$ of P and the expected value $E(P(1-P)/N)$. In the case of O, the conditional expected value of O is

$$Q = p_0(1-P) + Pp_1 = p_0 + (p_1 - p_0)P,$$

the covariance of O and Z is $(p_1 - p_0)C_{ZP}$, the conditional variance of O is $N^{-1}Q(1-Q)$, and the unconditional variance of O is

$$\sigma^2(O) = (p_1 - p_0)^2\sigma^2(P) + E(N^{-1}Q(1-Q)).$$

The correlation ρ_{TZ} of T and Z can readily be expressed in terms of the correlation ρ_{PZ}. One has

$$\rho_{TZ} = \frac{\rho_{PZ}}{[1 + E(P(1-P)/N)/\sigma^2(P)]^{1/2}}.$$

In like manner, the correlation of O and Z is

$$\rho_{OZ} = \frac{\rho_{PZ}}{[1 + (p_1 - p_0)^{-2} E(Q(1-Q)/N)/\sigma^2(P)]^{1/2}}.$$

For a simple but fairly reasonable case, consider P with mean and standard deviation of 0.01, so that 1% of words is expected to have the error; let $p_1 = 0.8$, so that 80% of words are identified as having the error given that the error exists, and $p_2 = 0.001$, so that one of each thousand words without the error is flagged as containing the error. Then $Q = 0.001 + 0.799P$ has mean 0.00899. Given essay lengths averaging around 300 words and $E(N^{-1})E(N)$ exceeds 1 (Cramér, 1946, p. 88), one might consider $E(P/N) = 0.00005$, so that $E(P/N)$ is $E(P)$ divided by 200. In this case, $E(P(1-P)/N)$ is rather close to $E(P/N)$. It follows that ρ_{TZ} is about $0.82\rho_{PZ}$ and ρ_{OZ} is about $0.77\rho_{PZ}$. Thus the software issue has reduced the size of the correlation with Z by a relatively modest amount. Many other scenarios can be considered with differing results. The main point is that imperfections in natural language processing need not necessarily have major impact in the use of the essay features under study for purposes of assessment.

Number of Discourse Units

In e-rater version 2.0, the number of discourse units in the essay is computed by an analysis of transitions between background, main points, supporting points, and conclusions (Attali et al., 2003; Burstein et al., in press). This feature of e-rater does not appear to be found in other methods for automated essay assessment. This number of discourse units provides a measure of the extent to which the essay has been developed. In the regression analysis of erater, the maximum of 0 and the number of units minus 8 is used as a predictor. An equivalent predictor in a linear regression is the maximum of eight and the number of discourse units. The use of 8 is based on the number of discourse units in a standard five-paragraph essay. In principle, the style program counts the number of paragraphs in an essay, but the conventions expected by style in its identification of paragraphs are not usually followed in essays scored by erater, so that style does not function correctly for the data under study. The number of paragraphs has been considered in assessment (Slotnick, 1972).

Once again, appropriate transformations need consideration. To examine this essay feature, distributional and correlational properties were examined for the original number of discourse units, the logarithm of the original, the square root of the original, and the minimum of 8 and the original number. Distributional properties for the various variables suggested no unusually unfavorable features with any transformation, but the logarithmic transformation appeared relatively undesirable, for the estimated skewness coefficients ranged from -0.928 for the first TOEFL essay to -1.304 for the second GMAT essay, and estimated kurtosis coefficients ranged from 0.577 for the first TOEFL essay to 1.798 for the second GMAT essay. The truncation transformation suffered to some degree from negative skewness, as might be expected from such a transformation. Estimated skewness coefficients ranged from -0.439 for the first TOEFL essay to -1.033 for the first GMAT essay. Results for the original variable and for the square root transformation were comparable on the whole. On the one hand, all transformations correlated highly with average holistic score. On the other hand, the estimated partial correlations with average holistic score given square root of number of characters were very small. The estimated correlations with average holistic score were about 0.5 for all essays and transformations. The estimated partial correlations never exceeded 0.100.

It should be noted that the computation of number of discourse units need not correspond to the calculation provided by expert human raters. As in the case of measures of grammar, usage, mechanics, and style, the relationship of errors to particular groups of subjects is a concern, and errors may reduce the correlation of number of discourse units to other variables of interest. Whether the reduction is substantial or even exists is unclear.

For completeness, a perl program was used to attempt to compute the number of paragraphs for essay and compare use of number of paragraphs to number of discourse units. This attempt did not demonstrate any advantage in use of number of paragraphs.

Average Length of Discourse Unit

The average length in words of a discourse unit is simply the ratio of the number of words to the number of discourse units. This ratio is used in the regression in e-rater version 2.0. It does not appear to have a close analogue in other attempts at essay assessment,

although average paragraph length has been considered (Slotnick, 1972). The use of transformations appears somewhat important here, for the kurtosis estimates are quite large for the original scale. For the four essays, the smallest estimate is 13.599. In addition, the smallest estimated skewness is 2.857. The most attractive simple transformation is minus the inverse square root. In this case, estimated skewness coefficients for the four prompts range from 0.027 to 0.071, while estimated kurtosis coefficients range from 0.227 to 0.617. The difficulty encountered with the distribution is that the number of words per unit is rather large if the writer does not use the transition expressions required for the program to identify discourse units. There is a modest correlation with average holistic score for both the original scale and the minus inverse square root transformation; however, the estimated partial correlation with average holistic score is negligible given the square root of the number of characters. Raw correlations estimates for the minus inverse square root range from 0.224 to 0.291 for the four prompts, but partial correlation estimates range from -0.042 to -0.094. Thus average number of words per unit is not an important predictor of average holistic score.

Standard Frequency Index

Use of relatively infrequent words might be an indicator of sophisticated writing (Finn, 1977). Investigation of such usage may be based on a standard frequency index (SFI) (Breland, Jones, & Jenkins, 1994; Breland, 1996; Breland & Jenkins, 1997, Carroll, 1970), a measure of word frequency on a logarithmic scale in which lower numbers indicate less frequent words. A value of 40 corresponds to a frequency of one per million. A value of 90 corresponds to a frequency of one in ten. Decreasing SFI corresponds to greater word difficulty. Consequently, it might be the case that lower values of the SFI would correspond to more sophisticated writing. Various implementations of the SFI have been used. A relatively recent one is the Breland index (Breland et al., 1994; Breland, 1996; Breland & Jenkins, 1997). This index is used in this report and in e-rater version 2.0.

Use of the SFI involves some basic decisions on summarization of data, for almost every word in an essay has a specific SFI found in the list of 179,195 words. The e-rater regression uses the fifth lowest SFI for a word in the essay, although the software also summarizes

the distribution of the SFI for words in the essay with an available SFI by use of the four smallest SFI values for words in the essay, the median SFI, and the tenth, twentieth, thirtieth, and fortieth percentiles of the SFI. For each essay, this report generally uses the median SFI of those words in the essay for which an SFI is recorded in a list of 179,195 words. In e-rater version 2.0, for each essay, the fifth lowest SFI value is used for essay words in the list of 179,195 words with an SFI. For this report, the author also computed the mean and standard deviation of the SFI for each essay.

Of the summary measures explored, the median SFI was the most attractive in practice. Results for this measure are summarized in Tables 20, 21, and 22. These tables suggest that the median SFI has a moderate deviation from a normal distribution and is moderately successful as a predictor of holistic score. Results are least satisfactory for the first GMAT prompt.

To some extent, this superior performance of the median over standard measures such as the mean and standard deviation may reflect the relative stability of the median in the presence of outliers. Such outliers may arise due to some technical issues that can affect use of the Breland SFI. Erroneously spelled words may occasionally appear in the word list as rare words. For instance, "acheived" appears in the list with an SFI of 20.3. In addition, errors in spelling may result in actual rare words. For example, "teh" may appear in an essay as a typographical error. This error is a word cited in the online version of the Oxford English Dictionary as a possible spelling of some concepts in Taoism or in Confucianism. The erroneous word has an SFI of 27. The distribution of the SFI may also be affected by the prompt. An added issue that may affect all summary measures is that a prompt that is more technical in nature may lead to use of less frequent words, and prompts that lead to use of proper nouns may also affect the SFI (Finn, 1977).

The current use of the fifth lowest SFI in the e-rater regression appears problematic. Results for the fifth lowest SFI are much less satisfactory in important respects. Estimated coefficients of skewness range from -1.110 to -2.226 and estimated coefficients of kurtosis range from 0.907 to 7.782. To be sure, the raw correlation of the fifth lowest SFI with average holistic score is more negative than the raw correlation of median SFI with average holistic score. The sample correlations range from -0.446 to -0.296. It should be

Table 20.
Summary Statistics for Median SFI

Program	Prompt	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	5183	58.000	1.419	0.800	1.230
GMAT	2	5158	60.148	1.941	0.171	0.606
TOEFL	1	4895	60.418	2.024	0.297	-0.442
TOEFL	2	4884	61.624	2.214	0.172	-0.126

Table 21.
Correlations for Median SFI

			Correlation		Corrected
Program	Prompt	Count	AHS	SCH	with AHS
GMAT	1	5183	-0.223	-0.155	-0.241
GMAT	2	5158	-0.307	-0.190	-0.333
TOEFL	1	4895	-0.327	-0.207	-0.350
TOEFL	2	4884	-0.329	-0.203	-0.352

Table 22.
Partial Correlation of SFI and Average Holistic Score Given Square Root of Number of Characters

			Partial correlation	
Program	Prompt	Count	Sample	Corrected
GMAT	1	5183	-0.169	-0.224
GMAT	2	5158	-0.284	-0.437
TOEFL	1	4895	-0.287	-0.378
TOEFL	2	4884	-0.293	-0.377

emphasized that this behavior reflects the natural correlation with essay length expected from any extreme value statistic. The sample partial correlations of the fifth lowest SFI and average holistic score given square root of number of characters range from -0.047 to 0.051. Thus median SFI appears to be the more reasonable measure to use. Nonetheless, it is worth noting that the actual effect on predictions from the regression analysis in e-rater that results from replacement of the fifth lowest SFI by the median SFI is very small.

As might be expected, median SFI and average word length are related. The sample correlations of median SFI and average word length range from -0.759 to -0.606 for the four prompts. Not surprisingly, the sample partial correlation of average word length and average holistic score given both square root of number of characters and median SFI ranges from -0.020 to 0.043.

Measures of Word Diversity

Variation in word choice can reasonably be regarded as a desirable feature of writing. A statistical complication to anticipate in the case of essays is that each word appears relatively infrequently. Consequently, it is desirable to employ measures that are well-behaved under such circumstances. A well-known measure that appears suitable is Simpson's index (Simpson, 1949; Gini, 1912). This measure considers the probability that two distinct randomly selected words from an essay are the same. Presumably it is desirable for such a measure to be small. If there are w distinct words with respective frequencies f_i for i from 1 to w, so that the essay has

$$n = \sum_{i=1}^{w} f_i$$

words, then Simpson's index for an essay is

$$[n(n-1)]^{-1} \sum_{i=1}^{w} f_i(f_i - 1).$$

Simpson's index is always between 0 and 1. If all words are the same, then the index is 1. If all words are distinct, then the index is 0. Variants on the index can also be considered in which stop words are omitted or in which stop words and prompt words are omitted.

Simpson's index does not appear to have been employed in essay assessment. In e-rater version 2.0, the ratio of types over tokens is employed to measure word diversity among content words. Here types is the number of distinct content words, and tokens is the total number of content words, so that the ratio never exceeds 1 and is never smaller than the reciprocal of the total number of content words. This ratio has the problem that it will normally decline as the number of words in the essay increases (Efron & Thisted, 1976). This behavior is not encountered with Simpson's index.

To illustrate results, the square root of Simpson's index and the types over tokens ratio are considered in Tables 23, 24, 25, 26, 27, and 28 with stop words removed. The square root is used due to the reduced magnitude of the coefficients of skewness and kurtosis relative to other simple transformations. The distributions of the statistics do not present major problems. Simpson's index has modest value as a predictor of average holistic score. The ratio of types to tokens is quite problematic, for the signs of correlations and partial correlations are different. As a consequence, use of Simpson's index appears more appropriate. Nonetheless, it should be noted, as in the case of the SFI, that a change from the ratio of types to tokens to Simpson's index has a quite small effect on predictions generated by the regression analysis in e-rater.

Table 23.
Summary Statistics for the Square Root of Simpson's Index

Program	Prompt	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	5183	0.1092	0.0237	0.480	0.966
GMAT	2	5158	0.0969	0.0224	0.870	1.471
TOEFL	1	4895	0.1012	0.0249	0.590	0.593
TOEFL	2	4884	0.1180	0.0279	0.604	0.586

Selection of Specific Words

Presumably the content of the essay should be appropriate for the prompt. Measurement of content is far from a trivial exercise. A simple approach is to consider the relationship of human holistic scores to the relative frequencies of specific words in an essay.

Table 24.
Correlations for Square Root of Simpson's Index

Program	Prompt	Count	Correlation AHS	Correlation SCH	Corrected with AHS
GMAT	1	5183	-0.114	-0.088	-0.123
GMAT	2	5158	-0.359	-0.310	-0.390
TOEFL	1	4895	-0.328	-0.235	-0.351
TOEFL	2	4884	-0.338	-0.295	-0.362

Table 25.
Partial Correlation of Square Root of Simpson's Index and Average Holistic Score Given Square Root of Number of Characters

Program	Prompt	Count	Partial correlation Sample	Partial correlation Corrected
GMAT	1	5183	-0.073	-0.097
GMAT	2	5158	-0.190	-0.291
TOEFL	1	4895	-0.248	-0.326
TOEFL	2	4884	-0.175	-0.225

Table 26.
Summary Statistics for the Types-to-tokens Ratio

Program	Prompt	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	5183	0.594	0.092	0.306	0.314
GMAT	2	5158	0.630	0.083	0.093	0.028
TOEFL	1	4895	0.652	0.095	-0.109	0.023
TOEFL	2	4884	0.595	0.093	0.098	0.007

For example, consider the word "east" in the first GMAT prompt. The sample correlation is about 0.29 between average holistic score and the square root of the ratio of the number of uses of "east" in the essay and the number of content words in the essay. For the first prompt, this sample correlation exceeds the sample correlation of average holistic score with such variables as median SFI and square root of the rate per word of grammar and usage errors. The challenge here is that the number of possible specific words to examine is very large, and it is not obvious in advance which specific words lead to relationships with

Table 27.
Correlations for Ratio of Types to Tokens

Program	Prompt	Count	Correlation AHS	Correlation SCH	Corrected with AHS
GMAT	1	5183	-0.468	-0.582	-0.506
GMAT	2	5158	-0.242	-0.401	-0.263
TOEFL	1	4895	-0.190	-0.386	-0.203
TOEFL	2	4884	-0.151	-0.307	-0.161

Table 28.
Partial Correlation of Ratio of Types to Tokens and Average Holistic Score Given Square Root of Number of Characters

Program	Prompt	Count	Partial correlation Sample	Partial correlation Corrected
GMAT	1	5183	0.015	0.020
GMAT	2	5158	0.218	0.334
TOEFL	1	4895	0.261	0.343
TOEFL	2	4884	0.194	0.249

holistic scores. For the first prompt, about 20,000 different words are encountered among all the essays under study. As a consequence, there is a substantial question of how to proceed.

One relatively simple approach relies on ordinary linear regression analysis for a specific prompt. The analysis considers prediction of a function H of holistic scores by use of relative word frequencies. In this section H will be average holistic score. To develop predictors based on relative word frequencies, for any word W, let $F(W)$ be the frequency of W in the essay. Let K be a positive integer, and let W_k, $1 \leq k \leq K$, be a list of content words, so that the W_k do not appear on a list of stop words. For example, for the first GMAT prompt, W_1 might be the word "service" and $F(W_1)$ might be the number of times the word "service" appears in the essay. Let M denote the total number of content words in the essay, so that M is the token count provided in e-rater for content words. Let G_k be the square root of the relative frequency $F(W_k)/M$. The G_k, $1 \leq k \leq K$, and other numerical essay features X_j, $1 \leq j \leq J$, related to holistic score are used to predict H. In this section J is 6, X_1 is the square root of the number of characters, X_2 is median SFI, X_3 is the square root of the rate per word of errors in grammar and usage, X_4 is the square

root of the rate per word of spelling errors, X_5 is the square root of the rate per word of other mechanical errors, and X_6 is the square root of Simpson's index for content words in the essay. The value of the content will be the change in the prediction of H that results from use of both X_j, $1 \leq j \leq J$, and G_k, $1 \leq k \leq K$, rather than from X_j, $1 \leq j \leq J$, alone.

The square root of the relative word frequency $F(W_k)/M$ is used rather than the relative frequency itself for two basic reasons. The transformation reduces skewness and kurtosis and it results in better predictor of H.

In an ideal world, the X_j, $1 \leq j \leq J$, and G_k, $1 \leq k \leq K$, are random variables with known finite means and variances and a known positive-definite covariance matrix, H is a random variable with known mean $E(H)$ and variance $\sigma^2(H)$, the covariance of X_j and H is known for $1 \leq j \leq J$, and the covariance of G_k and H is known for $1 \leq k \leq K$. In this ideal case, there is a uniquely defined best linear predictor

$$H_{XG} = E(H) + \sum_{j=1}^{J} \beta_{jXG}[X_j - E(X_j)] + \sum_{k=1}^{K} \gamma_k[G_k - E(G_k)]$$

of H in the sense that H_{XG} minimizes the expected squared error $E([H - H_{XG}]^2)$ among affine functions of X_j, $1 \leq j \leq J$, and G_k, $1 \leq k \leq K$. The constants β_{jXG} and γ_k are then known. In like manner, there is a best linear predictor

$$H_X = E(H) + \sum_{j=1}^{J} \beta_{jX}[X_j - E(X_j)]$$

that minimizes the mean squared error $E([H - H_X]^2)$ among affine functions of X_j, $1 \leq j \leq J$. The desired measure

$$C = H_{XG} - H_X = \sum_{j=1}^{J} (\beta_{jXG} - \beta_{jX})[X_j - E(X_j)] + \sum_{k=1}^{K} \gamma_k[G_k - E(G_k)]$$

is then readily computed for any given essay, and its properties are easily studied. The expected value $E(C) = 0$, the variance $\sigma^2(C)$ is the difference between the variance of H_{XG} and the variance of H_X, and the correlation of C and H is $\sigma(C)/\sigma(H) = (\rho^2_{HXG} - \rho^2_{HX})^{1/2}$, where ρ^2_{HX} is the coefficient of determination for prediction of H by the X_j and ρ^2_{HXG} is the coefficient of determination for prediction of H by the X_j and G_k.

In practice, a sample of n essays i, $1 \leq i \leq n$, is available. From the sample of essays, measurements H_i of H, X_{ij} for X_j, $1 \leq j \leq J$, and G_{ik} for G_k, $1 \leq k \leq K$, may be used to

estimate means, covariances, variances, and regression coefficients. The value H_{iXG} of H_{XG} for essay i has a least-squares approximation $\hat{H}_{iXG}$, and the value H_{iX} of H_X for essay i has least-squares approximation $\hat{H}_{iX}$. Thus the value C_i of C for essay i is approximated by $\hat{C}_i = H_{iXG} - \hat{H}_{iX}$. The observations $\hat{C}_i$ are not distributed as independent random variables, and the moments of the $\hat{C}_i$ differ from those of C. The issue becomes particularly significant if the number K of words is very large, for $\hat{H}_{iXG}$ will be H_i if K is sufficiently large. Given this consideration, for a given prompt, a relatively arbitrary decision was made for each prompt to consider those words W_k with a sample mean $\bar{F}_k$ of at least 0.15. This choice resulted, for the four prompts under study, in values of K of 172, 174, 127, and 138, respectively.

This approach to selection of K and W_k for a given prompt involves some complication in that K and W_k depend on the particular sample of essays; however, the selection of W_k has no direct relationship to holistic scores. In addition, if the essays can be regarded as a simple random sample from a hypothetical infinite population of essays and if no content word W exists with an expected frequency $E(F(W))$ equal to 0.15, then the law of large numbers can be applied to show that, as the sample size increases, the probability approaches 1 that K is the number of words W with $E(F(W)) > 0.15$ and the W_k are the content words with $E(F(W_k)) > 0.15$.

The conventional summary statistics for the $\hat{C}_i$, $1 \leq i \leq n$, are readily computed; however, some caution is required in terms of the relationship between the summary statistics and the corresponding population measures for C. Under random sample, the $\hat{C}_i$ all have a common distribution, that of a random variable that may be denoted by $\hat{C}$. Standard properties of linear regression show that the sample mean of the $\hat{C}_i$ is always equal to 0, the expected value $E(C)$ of C (Draper & Smith, 1998, p. 61). Thus $\hat{C}$ has expectation 0. The sample variance $s^2(\hat{C})$ of the $\hat{C}_i$ is

$$s^2(\hat{C}) = (n-1)^{-1} \sum_{i=1}^{n} \hat{C}_i^2,$$

so that the expectation of $s^2(\hat{C})$ is $[n/(n-1)]\sigma^2(\hat{C})$, where the variance $\sigma^2(\hat{C})$ of $\hat{C}$ is $E(\hat{C}^2)$. Thus the sample standard deviation $s(\hat{C})$ of the $\hat{C}_i$ estimates the standard deviation $\sigma(\hat{C})$ of $\hat{C}$.

The sample variance $s^2(\hat{C})$ also provides an estimate of the population variance $\sigma^2(C)$ of C; however, bias is a more significant issue in this case. Because

$$s^2(\hat{C}) = (n-1)^{-1} \sum_{i=1}^{n} (\hat{H}_{iXG} - \hat{H}_{iX})^2,$$

it follows that

$$s^2(\hat{C}) = \frac{K}{n-1} \, \mathrm{MSE}(G|X),$$

where $\mathrm{MSE}(G|X)$ is the mean squared error for testing the hypothesis that $\gamma_k = 0$ for $1 \leq k \leq K$ (Draper & Smith, 1998, chap. 6). The expected value of $s^2(\hat{C})$ under a traditional model for regression analysis in which $H - H_{XG}$ is independent of the X_j and G_k is exactly

$$\frac{n}{n-1}\sigma^2(C) + \frac{K}{n-1}\sigma^2(H - H_{XG})$$

if the X_j and W_k had a joint continuous distribution. In general, the expected value of $s^2(\hat{C})$ differs from $\sigma^2(C)$ by a term of order K/n (Rao, 1973, p. 222). For fixed K and large n, $s^2(\hat{C})$ does provide a consistent estimate of $\sigma^2(C)$ (i.e., the variance of $\hat{C}$ converges to the variance of C); however, the ratios K/n in the analysis under consideration are not small enough to rely on this result. Application of jackknifing to $[(n-1)/n]s^2(\hat{C})$ provides an estimate v for $\sigma^2(\hat{C})$ with a bias of order $1/n^2$ (Miller, 1964). The square root $v^{1/2}$ provides an alternative estimate of $\sigma(C)$ with a smaller bias problem. The estimate v can be obtained in a relatively straightforward fashion by use of standard computer packages such as SAS. One finds that

$$v = \frac{n-1}{n} s^2(\hat{C}) - \bar{U},$$

where $\bar{U}$ is the average of U_i for $1 \leq i \leq n$ and U_i is found by the following procedure. There are n by n symmetric matrices $\mathbf{R}_{XG}$ and $\mathbf{R}_X$ called hat matrices associated with the two regressions used to compute $\hat{C}$ (Draper & Smith, 1998, pp. 205–207). For i and i' from 1 to n, row i and column i' of $\mathbf{R}_{XG}$ is $R_{ii'XG}$, while row i and column i' of $\mathbf{R}_X$ is $R_{ii'X}$. The matrix $\mathbf{R}_{XG}$ is determined by the X_{ij}, $1 \leq j \leq J$, and W_{ik}, $1 \leq k \leq K$, for $1 \leq i \leq n$, while $\mathbf{R}_X$ is determined by the X_{ij}, $1 \leq\leq J$, $1 \leq i \leq n$,

$$\hat{H}_{iX} = \sum_{i'=1}^{n} R_{ii'X} H_{i'},$$

Table 29.
Summary Statistics for the Statistic $\hat{C}$

Program	Prompt	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	5183	0	0.382	-0.425	0.103
GMAT	2	5158	0	0.214	0.006	-0.161
TOEFL	1	4895	0	0.141	0.042	-0.061
TOEFL	2	4884	0	0.157	0.012	0.167

and

$$\hat{H}_{iXG} = \sum_{i'=1}^{n} R_{ii'XG} H_{i'}$$

(Draper & Smith, 1998, pp. 205–209). The R_{iiXG} and R_{iiX} are never negative or greater than 1. They are positive and less than 1 if the least-squares estimates $\hat{\beta}_j$ of β_j and $\hat{\gamma}_k$ are uniquely defined. Let r_{iXG} be the conventional residual $H_i - \hat{H}_{iXG}$, and let r_{iX} be the residual $H_i - \hat{H}_{iX}$ (Draper & Smith, 1998, p. 60). Then

$$U_i = \frac{r_{iXG}^2 R_{iiXG}}{1 - R_{iiXG}} - \frac{r_{iX}^2 R_{iiX}}{1 - R_{iiX}}.$$

In programs such as SAS, the R_{iiXG}, R_{iiX}, r_{iXG}, and r_{iX} are provided as optional output in the regression procedure.

Table 29 provides a basic look at the distribution of $\hat{C}$. This table does not suggest anything remarkable. Nonetheless, some difficulty is evident from the jackknifed estimates $v^{1/2}$ for $\sigma(C)$. These estimates for the four prompts are 0.372, 0.197, 0.105, and 0.119, respectively. Thus the sample standard deviations $s(\hat{C})$ appear to be somewhat higher than $v^{1/2}$, especially for the TOEFL prompts. Thus the effects of estimation of C_i by $\hat{C}_i$ appear to be of some significance.

The correlation of $\hat{C}_i$ and H_i is readily estimated. Elementary arguments show that the sample covariance of the $\hat{C}_i$ and H_i is

$$(n-1)^{-1} \sum_{i=1}^{n} \hat{C}_i H_i = s^2(\hat{C})$$

(Draper & Smith, 1998, p. 206), so that the sample correlation of $\hat{C}_i$ and H_i is $s(\hat{C})/s(H)$. Because the covariance of C and H is the variance of C (Rao, 1973, p. 267), the correlation

Table 30.
Estimated Correlations of Average Holistic Score H_i ***with*** $\hat{C}_i$ ***and*** C_i

			Sample correlation		Corrected
Test	Prompt	Count	for $\hat{C}$	for C	for C
GMAT	1	5183	0.342	0.332	0.359
GMAT	2	5158	0.201	0.185	0.201
TOEFL	1	4895	0.135	0.101	0.109
TOEFL	2	4884	0.154	0.117	0.125

of C and H may be estimated by $v/s(H)$. This ratio can be corrected for attenuation in the usual fashion. Thus results in Table 30 are obtained. These results suggest that the variable C has a modest correlation with average holistic score, with the relationship weakest for the TOEFL prompts.

An alternative correlation criterion can be used based on application to a new observation. The estimate $\hat{C}_i$ can be written

$$\hat{C}_i = \sum_{j=1}^{J} (\hat{\beta}_{jXG} - \hat{\beta}_{jX})(X_{ij} - \bar{X}_j) + \sum_{k=1}^{K} \hat{\gamma}_k (G_{ik} - \bar{G}_k),$$

where hats are used to denoted least-squares estimates and bars are used to denote sample means for the sample of n essays. Consider the correlation of H to the estimate

$$C^* = \sum_{j=1}^{J} (\hat{\beta}_{jXG} - \hat{\beta}_{jX})(X_j - \bar{X}_j) + \sum_{k=1}^{K} \hat{\gamma}_k (G_k - \bar{G}_k)$$

derived from the observed sample and applied to the random variables X_j and G_k. The variables C^* and H are not directly observed, but their correlation can be estimated by use of deleted observations. For each observation i, β_{jXG}, β_{jX}, γ_k, $E(X_j)$, and $E(G_k)$ are estimated from the sample of $n-1$ observations obtained by deleting observation i from the regression analysis. The resulting estimates may be denoted by insertion of a subscript (i), so that $\hat{\beta}_{jXG(i)}$ is the estimate of β_{jXG} obtained by deletion of observation i. The deleted approximation to C_i is then

$$\hat{C}_{(i)} = \sum_{j=1}^{J} (\hat{\beta}_{jXG(i)} - \hat{\beta}_{jX(i)})(X_j - \bar{X}_{j(i)}) + \sum_{k=1}^{K} \hat{\gamma}_{k(i)} (G_k - \bar{G}_{k(i)}).$$

Table 31.
Estimated Correlation of Average Holistic Score With C^*

Test	Prompt	Sample count	Corrected correlation	Correlation
GMAT	1	5183	0.323	0.349
GMAT	2	5158	0.170	0.185
TOEFL	1	4895	0.095	0.102
TOEFL	2	4884	0.112	0.120

The sample correlation of $\hat{C}_{(i)}$ and H_i may then be considered. One might intuitively expect that computation of $\hat{C}_{(i)}$ would be extremely cumbersome; however, the actual computation is quite simple, for

$$\hat{C}_{(i)} = r_{iXd} - r_{iXGd}$$

for the deleted residuals

$$r_{iXd} = \frac{r_{iX}}{1 - R_{iiX}}$$

and

$$r_{iXGd} = -\frac{r_{iXG}}{1 - R_{iXG}}$$

(Weisberg, 1985). The term deleted residual or PRESS residual is used because

$$r_{iXd} = H_i - \bar{H}_{(i)} - \sum_{j=1}^{J} \hat{\beta}_{jX(i)}(X_{ij} - \bar{X}_{j(i)})$$

is the difference between the observed score H_i and the prediction of H_i based on the observations other than i. The new sample correlations are listed in Table 31. The new results are relatively similar to the previous ones, so that conclusions are not changed.

The definition of C permits a simple analysis of the partial correlation given square root of number of characters (X_1). Let $\hat{\rho}$ be the sample correlation of X_1 and H, and let $s(H|X_1) = s(H)(1 - \hat{\rho}^2)^{1/2}$. Then the sample partial correlation of $\hat{C}_i$ and H_i given X_1 is the ratio $s(\hat{C})/s(H|X_1)$. One may approximate the partial correlation of C and H given X_1 by $v^{1/2}/s(H|X_1)$. Sample partial correlations can also be computed for $\hat{C}_{(i)}$ and H_i given X_{i1}. Results are shown in Table 32. As can be predicted given the definition of C, the partial correlations with holistic score are somewhat larger than are the original correlations. This effect is particularly important for the first GMAT prompt.

Table 32.
Estimated Partial Correlations of Average Holistic Score H_i
With $\hat{C}_i$, C_i, ***and*** $C_{(i)}$ ***Given Square Root of Number of Characters***

			Partial correlation		
Test	Prompt	Count	for $\hat{C}$	for C	for $C_{(i)}$
GMAT	1	5183	0.592	0.515	0.560
GMAT	2	5158	0.391	0.338	0.331
TOEFL	1	4895	0.246	0.178	0.176
TOEFL	2	4884	0.275	0.202	0.202

It should be noted that the overall prediction of holistic score from the combination of the X_j and G_k is rather impressive, especially if corrected for attenuation. In Table 33, conventional R^2 statistics are provided and corrected for attenuation for the basic features X_j and for the combination of the X_j and G_k. In addition, R^2 statistics are computed based on deleted residuals to consider the proportional reduction in squared error from application of the regression to new H, X_j, and G_k. As an example, the conventional R^2 for prediction by the X_j is

$$R^2_X = 1 - \frac{\sum_{i=1}^n r^2_{iX}}{\sum_{i=1}^n (H_i - \bar{H})^2}.$$

With deleted residuals, one obtains

$$R^2_{X*} = 1 - \frac{\sum_{i=1}^n r^2_{iXd}}{[n/(n-1)]^2 \sum_{i=1}^n (H_i - \bar{H})^2}.$$

Note that the ith deleted residual for a constant predictor is

$$H_i - \frac{n\bar{H} - H_i}{n-1} = \frac{n}{n-1}(H_i - \bar{H}).$$

For prediction of H by X_j and G_k, one has

$$R^2_{XG} = 1 - \frac{\sum_{i=1}^n r^2_{iXG}}{\sum_{i=1}^n (H_i - \bar{H})^2}.$$

With deleted residuals, one obtains

$$R^2_{XG*} = 1 - \frac{\sum_{i=1}^n r^2_{iXGd}}{[n/(n-1)]^2 \sum_{i=1}^n (H_i - \bar{H})^2}.$$

In all cases, there is an appreciable gain from addition of the G_k; however, the effect is most pronounced in the first GMAT prompt.

Table 33.
Estimated Coefficients of Determination for Prediction of Average Holistic Score

	GMAT		TOEFL	
Prompt	1	2	1	2
Count	5183	5158	4895	4884
R^2_X	0.695	0.782	0.780	0.758
Corrected R^2_X	0.810	0.922	0.893	0.866
R^2_{XG}	0.812	0.822	0.798	0.782
Corrected R^2_{XG}	0.947	0.970	0.914	0.893
R^2_{X*}	0.695	0.781	0.779	0.757
Corrected R^2_{X*}	0.810	0.921	0.893	0.865
R^2_{XG*}	0.798	0.809	0.787	0.768
Corrected R^2_{XG*}	0.931	0.954	0.901	0.877

The approach toward addition of the square roots of relative frequencies adopted here tends to minimize their apparent impact, for the linear combination used reflects the relationship with average holistic score after other variables have already been considered. If one simply uses a regression of average holistic score on the G_k, then the multiple correlation statistics for the four prompts are 0.791, 0.722, 0.663, and 0.655, respectively, so that the relationship of the relative frequencies to average holistic score is quite strong. An even more extreme result is obtained if square roots $F(W_k)/M]^{1/2}$ of relative frequencies are replaced by square roots $[F(W_k)]^{1/2}$ of frequencies, and the additional predictors square root of number of stop words and square root $M^{1/2}$ of content words other than W_k, $1 \leq k \leq K$, are added. This replacement yields multiple correlation statistics of 0.901, 0.897, 0.872, and 0.863, respectively. Use of absolute frequencies rather than relative frequencies can be interpreted as leading to greater predictive power because absolute frequencies reflect essay length. Nonetheless, it is certainly possible to interpret analysis as suggestive of a very high importance of content. Very similar results are obtained if for each frequency F, the square root $F^{1/2}$ is replaced by $\log(F+1)$.

Content vector analysis. The regression approach differs somewhat from the approach used in e-rater version 2.0. Here content vector analysis is employed (Burstein et al., in press; Attali et al., 2003). In this approach, for a given prompt, a collection of essays is

selected as a training set. The training set contains six groups, with essays in each group having the same resolved holistic score. Fifty essays are selected from each group for scores 2 to 6, and 15 essays are selected from the group of scores of 1. The lower number of scores of 1 reflects the lower frequency of a resolved holistic score of 1. The word distributions exclusive of stop words for individual essays are compared to the corresponding word distributions for the six groups. In this comparison, let there be L distinct words V_k, $1 \leq k \leq L$, other than stop words in the training set of essays. Let $v_k = F(V_k)/M$ be the fraction of words, other than stop words, in an individual essay that are equal to V_k, let u_{kh} be the fraction of content words in the training set for resolved holistic score h that are equal to V_k, and let t_k be the logarithm of the inverse of the fraction of training essays in which word V_k appears. Let $w_k = v_k t_k$ and $x_{kh} = u_{kh} t_k$ be the weighted word frequencies for word V_k in the essay and in group h. Note that the weights are 0 if the word appears in all essays in the training set. Then the cosine for holistic score h computed in e-rater version 2.0 is algebraically equivalent to

$$a_h = \frac{\sum_{k=1}^{L} w_k x_{kh}}{\left[\sum_{k=1}^{L} w_k^2\right]^{1/2} \left[\sum_{k=1}^{L} x_{kh}^2\right]^{1/2}}.$$

In version 2.0 of e-rater, the cosine a_6 and the integer τ such that $a_\tau \geq a_h$ for all h are used as predicting variables in the regression. As in the regression approach to essay content, a_6 and τ depend on essays other than the essay directly under study. The effects of the use of a training set are difficult to analyze. If one ignores this issue, some basic summary statistics and correlations can be examined, but it is difficult to ascertain the extent to which results are biased. The sample distributions of a_6 and τ appear unremarkable except for the moderately negative kurtosis of τ. The sample correlations of a_6 and τ with average holistic score are relatively large, and there is a substantial sample partial correlation of a_6 and τ with average holistic given square root of number of characters. Nonetheless, the relationship of a_6 and average holistic score is somewhat decreased given the square root of the number of characters. This result reflects a correlation of a_6 with square root of number of characters of about 0.6 for each prompt. Consider Tables‘34, 35, and 36. Corrections for attenuation are not considered due to the use of the training set.

If the previously used variables X_j, $1 \leq j \leq J$, and a_6 and τ are used to predict average

Table 34.
Summary Statistics for Variables in Content Vector Analysis

Program	Prompt	Function	Count	Average	Standard deviation	Skewness	Kurtosis
GMAT	1	a_6	5183	0.224	0.062	0.082	-0.276
GMAT	1	τ	5183	4.139	1.324	-0.146	-1.036
GMAT	2	a_6	5158	0.214	0.051	0.036	-0.124
GMAT	2	τ	5183	4.274	1.282	-0.149	-1.103
TOEFL	1	a_6	4895	0.201	0.046	-0.192	0.070
TOEFL	1	τ	4895	4.497	1.177	-0.145	-1.072
TOEFL	2	a_6	4884	0.212	0.051	-0.108	-0.005
TOEFL	2	τ	4884	4.481	1.192	-0.154	-1.034

Table 35.
Correlations of Variables in Content Vector Analysis With Average Holistic Score

			Correlation	
Program	Prompt	Count	a_6	τ
GMAT	1	5183	0.702	0.574
GMAT	2	5158	0.617	0.536
TOEFL	1	4895	0.626	0.556
TOEFL	2	4884	0.581	0.499

Table 36.
Partial Correlations of Variables in Content Vector Analysis and Average Holistic Score

			Partial correlation	
Program	Prompt	Count	a_6	τ
GMAT	1	5183	0.450	0.560
GMAT	2	5158	0.203	0.470
TOEFL	1	4895	0.231	0.431
TOEFL	2	4884	0.184	0.398

holistic score, then the resulting R^2 statistics for the four prompts are 0.788, 0.812, 0.789, and 0.774, respectively. Use of the regression from e-rater version 2.0 results in slightly larger values of R^2. Respective results for the four prompts are 0.791, 0.817, 0.796, and 0.784. The results for these two regressions are rather similar to those obtained with the combination of X_j and G_k, as evident from Table 33. A fully satisfactory comparison requires some adjustment for the effect of the training set. It is not clear how to accomplish this step in a reasonable fashion.

It should be noted that the prediction of average holistic score yields a somewhat higher R^2 statistic than does prediction of resolved holistic score, even though the variables from content vector analysis were derived by a training set defined by means of resolved holistic score. For each of the four prompts under study, the reduction in conventional R^2 for the regression with the X_j, a_6, and τ is at least 0.046 if resolved holistic score is predicted rather than average holistic score.

A reasonable case can be made that the regression analysis with square root of relative frequencies is preferable to the content vector analysis to the extent that its statistical properties are more readily investigated and the predictive properties appear quite similar. On the other hand, there is no reason to assume that a substantial improvement in predictions of holistic scores will result from use of square roots of relative frequencies.

Latent semantic analysis. In the Intelligent Essay Assessor™, word choice is examined by latent semantic analysis (Landauer et al., 1998; Landauer et al., 2003). In this analysis, one may let L' be the number of words V'_k that appear in at least two essays. Let F_{ik} be the number of words in essay i equal to V'_k, let F_{+k} be the sum of F_{ik} for all essays i, and let $f_{ik} = F_{ik}/F_{+k}$. A singular value decomposition is performed on the matrix with row i and column k equal to

$$S_{ik} = \frac{\log[F(V'_k) + 1]}{-\sum_{i=1}^{n} f_{ik} \log f_{ik}}.$$

Here the convention is used that $0 \log 0 = 0$. The singular value decomposition is employed to provide a smoothed approximation to the S_{ik}. These smoothed approximations are then used to evaluate similarity of essays to each other or to given passages. A variety of techniques are employed to assign holistic essay scores. Some approaches employ training sets in a manner similar to that used in content vector analysis.

Like content vector analysis, latent semantic analysis is very difficult to analyze. The very high dimensions involved in the singular value decomposition and the use of training sets are both problems. One issue worth noting is that the definition of S_{ik} depends on the essay length, so that it is reasonable to expect that H_i can be predicted with considerable accuracy by the values S_{ik} used as a basis for latent semantic analysis.

2 Variation by Prompt and Population

Variations of features by population and by prompt within population can be examined to some degree by use of the GMAT and TOEFL prompts under study. This analysis is limited due to the lack of common prompts. In addition, the number of available prompts is limited. Nonetheless, some indications can be obtained for variability of features with respect to prompts. The reader should regard any results for holistic scores with particular caution given possible use of program procedures to control the distribution of holistic scores for specific prompts. The basic analysis simply examines the R^2 statistics for prediction of essay features based on knowledge of the prompt or program. For a summary of findings, see Table 37. The abbreviations for grammar, usage, mechanics, and style features follow those in Table 16. The Simpson's index is for all words except stop words. The variable $\hat{C}$ is not examined because its sample mean for each prompt must be 0. Some caution is required with a_6 and τ due to the use of a training set.

Table 37.
R^2 Statistics for Prediction of Essay Features From Prompt

Feature	Prompts	Programs
Average holistic score	0.017	0.017
Resolved holistic score	0.014	0.011
Root no. chars.	0.140	0.139
Root GU	0.041	0.039
Root spell	0.050	0.030
Root OM	0.043	0.044
Med. SFI	0.319	0.176
Root Simpson's index	0.094	0.016
a_6	0.023	0.013
τ	0.016	0.013

For the most part, association of result with prompt or program is fairly weak. The main exceptions involve word choice, both in terms of the SFI and in terms of average word length. To some degree, it is evident from Tables 8 and 20 that TOEFL examinees have some tendency to use more common and shorter words that GMAT examinees; however, there are also evident variations by prompt.

The modest associations may still be important if the features are used in an assessment to evaluate aspects of writing. Especially in the case of median SFI, interpretation of results would appear to require adjustment for the prompt and program involved. This issue also arises with the square root of the number of characters. As evident from Table 6, the sample means for the GMAT prompts are substantially larger than for the TOEFL prompts, although the variations for prompts within programs appear to be quite small.

The analysis here does not address the possible use of common regression equations in e-rater version 2.0 for classes of prompts (Attali et al., 2003), for differences in the behavior of predictors do not affect appropriateness of regression coefficients. To consider this issue, regard all essays from the four prompts as a single sample. Let average holistic score be the dependent variable. If a regression uses X_j, $1 \leq j \leq J$, as independent variables without regard to prompt, then the R^2 is only 0.544. If a separate regression equation is used for each program, then R^2 is 0.752, while a separate regression for each prompt yields an R^2 of 0.756. It is interesting to note that a regression model for all four prompts with a dummy variable for prompt and with the six X_j yields an R^2 of 0.755. If the dummy variable for prompt is replaced by a dummy variable for program, then R^2 is still 0.751. Thus maintaining the same regression coefficients associated with the variables X_j for each prompt does not cause a major deterioration of predictive power, although the intercept of the regression equation appears to require an adjustment by program.

Some added information on variability by prompt and population can be derived from essays obtained via Criterion from a tenth grade in Florida (Shermis et al., 2004). The data are limited in terms of the number of students involved, and the sample was from a random sample of American students. There are seven different prompts and 562 students in the study; however, not all students respond to all prompts, and some students have more than one response to a prompt. In addition, the data are obtained under less controlled

conditions than are present in GMAT and TOEFL tests, and the student motivation may not be the same as in a high-stakes test. The first response was the only one used in the analysis presented here. With the restriction that only essays of at least 25 words are used, there are 2,595 essays examined. Thus the data are somewhat more limited than for the GMAT and TOEFL prompts. A further restriction is that resolved holistic score are available to the author, but ratings from multiple judges are not available. Table 38 provides an overall summary of means for GMAT essays, TOEFL essays and Criterion^{SM} essays. The differences in means are in predictable directions, for the students in high school have average measures of errors in grammar, usage, and mechanics that are higher than for the other groups; the average of the median SFI is higher than in other groups; and the square root of the Simpson index is also higher on average for the Criterion essays.

Table 38.
Means of Essay Features for Essays
From GMAT, TOEFL, and Criterion

Feature	GMAT	TOEFL	Criterion
Resolved holistic score	4.0390	4.299	3.905
Root no. chars.	37.4920	32.360	34.503
Root GU	0.0540	0.075	0.094
Root spell	0.1150	0.139	0.147
Root OM	0.0330	0.054	0.111
Med. SFI	59.0710	61.020	62.251
Root Simpson's index	0.1030	0.110	0.119
a_6	0.2193	0.206	0.148
τ	4.2060	4.489	4.118

Variations of essay features by prompt may also be examined for the high school essays. Table 39 summarizes results. It may be compared to Table 37. The tables are fairly similar, although the variables from content vector analysis appear more dependent on the prompt in the data from Florida.

3 Variation by Individual

The data from the Florida high school permits some examination of reliability, although analysis is a bit more complicated due to the variation in the number of essays each student

Table 39.
R^2 Statistics for Prediction of Essay Features From Prompt for Criterion Data

Feature	R^2
Resolved holistic score	0.135
Root no. chars.	0.133
Root GU	0.009
Root spell	0.012
Root OM	0.061
Med. SFI	0.210
Root Simpson's index	0.237
a_6	0.263
τ	0.126

writes. Thus the usual textbook approach to generalizability must be modified as in common approaches to analysis of variance (Brennan, 1992; Scheffé, 1959). For each essay feature in Table 39, an analysis of variance is used in which prompt, examinee, and prompt by examinee interaction are considered to be independent random effects with respective variances σ_p^2, σ_i^2, and σ^2. Thus a single essay feature Y has the decomposition

$$Y = \mu + P + I + \epsilon,$$

where P, the prompt effect has mean 0 and variance $\sigma_p^2(Y)$, I, the examinee effect, has mean 0 and variance $\sigma_i^2(Y)$, and ϵ, the prompt by examinee interaction, has mean 0 and variance $\sigma^2(Y)$. It is assumed that P, I, and ϵ are uncorrelated. Implicit in this model are hypothetical infinite populations of prompts and examinees. For an observed prompt p and examinee i such that examinee i writes an essay on prompt p, the value of Y is Y_{pi}. The observed prompts may be numbered from 1 to u, and the individuals may be numbered from 1 to $n = 562$. The indicator χ_{pi} is 1 if examinee i responds to prompt p, and χ_{pi} is otherwise 0. The value of Y_{pi} is arbitrarily set to 0 if χ_{pi} is 0. There are $n_{p+} > 0$ individuals responding to prompt p, and $n_{+i} > 0$ essays written by individual i. The average feature value Y for prompt p is then

$$\bar{Y}_{p\cdot} = \frac{1}{n_{p+}} \sum_{i=1}^{n} \chi_{pi} Y_{pi},$$

and the average feature value for examinee i of Y is

$$\bar{Y}_{\cdot i} = \frac{1}{n_{+i}} \sum_{p=1}^{u} \chi_{pi} Y_{pi}.$$

Let N be the total number of essays. Let P have value P_p for prompt p, let I have value I_i for examinee i, and let ϵ have value ϵ_{pi} for prompt p and examinee i. Then

$$\bar{Y}_{p\cdot} = \mu + P_p + \frac{1}{n_{p+}} \sum_{i=1}^{n} \chi_{pi}(I_i + \epsilon_{pi}),$$

and

$$\bar{Y}_{\cdot i} = \mu + I_i + \frac{1}{n_{+i}} \sum_{p=1}^{u} \chi_{pi}(P_p + \epsilon_{pi}).$$

Let

$$\mathrm{SSE}(Y|P) = \sum_{p=1}^{u} \sum_{i=1}^{n} \chi_{pi}(Y_{pi} - \bar{Y}_{p\cdot})^2$$

be the residual sum of squares and let

$$\mathrm{MSE}(Y|P) = (N-u)^{-1}\,\mathrm{SSE}(Y|P)$$

be the residual mean square for the hypothesis that $\sigma_i^2(Y) = 0$. Then $\mathrm{MSE}(Y|P)$ has expected value $\sigma_i^2(Y) + \sigma^2(Y)$. In like manner, let

$$\mathrm{SSE}(Y|I) = \sum_{i=1}^{n} \sum_{p=1}^{u} \chi_{pi}(Y_{pi} - \bar{Y}_{\cdot i})^2$$

be the residual sum of squares and let

$$\mathrm{MSE}(Y|I) = (N-n)^{-1}\,\mathrm{SSE}(Y|I)$$

be the residual mean square for the hypothesis that $\sigma_p^2(Y) = 0$. Then $\mathrm{MSE}(Y|I)$ has expected value $\sigma_p^2(Y) + \sigma^2(Y)$. Let $\mathrm{SSE}(Y|PI)$ be the residual mean square for the hypothesis that, conditional on the observed examinees and prompts, an additive model for the observed Y_{pi} holds, so that $\mathrm{SSE}(Y|PI)$ is the minimum of

$$\sum_{p=1}^{u} \sum_{i=1}^{n} \chi_{pi}(Y_{pi} - a - b_p - c_i)^2$$

for real a, b_p, and c_i such that $\sum_{p=1}^{u} b_u = \sum_{i=1}^{n} c_i = 0$. Given that the a, b_p, and c_i are uniquely identified, the residual mean square $\mathrm{MSE}(Y|PI)$ for the additive model is

$$\mathrm{MSE}(Y|PI) = (N - n - u + 1)^{-1}\,\mathrm{SSE}(Y|PI).$$

The expectation of $s^2(Y) = \mathrm{MSE}(Y|PI)$ is σ^2. Thus $\sigma_p^2(Y)$ has an unbiased estimate

$$s_p^2(Y) = \mathrm{MSE}(Y|I) - \mathrm{MSE}(Y|PI),$$

and $\sigma_i^2(Y)$ has an unbiased estimate

$$s_i^2(Y) = \mathrm{MSE}(Y|P) - \mathrm{MSE}(Y|PI).$$

Under the model, an examinee with examinee effect I has expected feature value Y of $\mu + I$ in response to a randomly selected prompt. The value $T = \mu + I$ may be regarded as the true score of I on feature Y. The variance of T is $\sigma_i^2(Y)$. Consider an approximation to T based on responses of the individual to r randomly selected prompts. If the approximation used is the average value $\bar{Y}$ of Y observed for the r prompts, then $\bar{Y} - T$ has mean 0, is uncorrelated with T, and has variance $r^{-1}[\sigma_p^2(Y) + \sigma^2(Y)]$. The reliability observed for the average is

$$\rho_{1r}^2(Y) = \frac{\sigma_i^2(Y)}{\sigma_i^2(Y) + r^{-1}[\sigma_p^2(Y) + \sigma^2(Y)]}.$$

On the other hand, suppose that the values of P are known for each prompt. Then the average $\bar{P}$ for the sampled prompts is available, and $\bar{Y}$ may be replaced by $\bar{Y} - \bar{P}$. In this case, ($\bar{Y} - \bar{P} - T$ has mean 0, is uncorrelated with T, and has variance $\sigma_i^2(Y) + r^{-1}\sigma^2(Y)$), so that the reliability is now

$$\rho_{2r}^2(Y) = \frac{\sigma_i^2(Y)}{\sigma_i^2(Y) + r^{-1}[\sigma_p^2(Y) + \sigma^2(Y)]}.$$

Although $\bar{P}$ will not really be known, the error in estimating $\bar{P}$ will be negligible in cases in which a large number of subjects responds to each essay.

Reliability coefficients are estimated by substitution of the unbiased estimates for the population quantities. For example, $\rho_{2r}^2(Y)$ has estimate

$$\hat{\rho}_{2r}^2(Y) = \frac{s_i^2(Y)}{s_i^2(Y) + r^{-1}s^2(Y)}.$$

Table 40.
Variability of Features by Prompt and Individual

Feature	$\hat{\rho}^2_{11}(Y)$	$\hat{\rho}^2_{21}(Y)$	$\hat{\rho}^2_{12}(Y)$	$\hat{\rho}^2_{22}(Y)$	$\hat{\rho}^2_{14}(Y)$	$\hat{\rho}^2_{24}(Y)$
Res. hol. score	0.449	0.520	0.620	0.684	0.766	0.813
Root no. chars.	0.532	0.601	0.694	0.751	0.820	0.858
Median SFI	0.265	0.344	0.419	0.511	0.591	0.677
Root GU	0.290	0.292	0.450	0.452	0.620	0.623
Root spelling	0.470	0.477	0.640	0.646	0.780	0.785
Root OM	0.529	0.563	0.692	0.721	0.818	0.838
Root Simpson's index	0.286	0.382	0.445	0.553	0.616	0.712
a_6	0.206	0.294	0.342	0.454	0.510	0.625
τ	0.147	0.170	0.256	0.291	0.408	0.451

Results are summarized in Table 40. Results for variables from content vector analysis should be approached with caution due to the use of a training set. Use of the $\hat{C}$ statistic does not appear practical for these data due to the small sample sizes for each prompt.

These results suggest that reliability of individual essay features is modest, especially in the case of grammar, usage, and the variables from content vector analysis. Nonetheless, use of four prompts does provide fairly reliable measurement for resolved holistic score if adjustment is made for prompt. It should be noted that the observed reliability would not suffice even then for a high stakes test. Reliability for four prompts is also fairly satisfactory for square root of number of characters, square root of rate of other errors in mechanics, and square root of rate of spelling errors. The reliability results for resolved holistic score do not appear to be unusual relative to reliability results previously reported (Breland et al., 1999).

4 Combinations

It is obviously possible for linear combinations of features to be more reliable than individual features. The limit of this concept can be explored by use of multivariate analysis of variance. Consider variables X_j, $1 \leq j \leq J$ and a new linear combination

$$Y = \mathbf{c}'\mathbf{X} = \sum_{j=1}^{J} c_j X_j.$$

Recall the notation in Section 3. Consider maximization of $\hat{\rho}_{2r}^2(Y)$ by optimal choice of c_j. This maximization corresponds to maximization of

$$\frac{\mathrm{MSE}(Y|P) - \mathrm{MSE}(Y|PI)}{\mathrm{MSE}(Y|P) - \mathrm{MSE}(Y|PI) + r^{-1}\,\mathrm{MSE}(Y|PI)}.$$

Equivalently, $\mathrm{MSE}(Y|P)/\,\mathrm{MSE}(Y|PI)$ is maximized without regard to the number r of essays. Recall from standard analysis of variance that the mean square for the examinee effect of Y given the prompt effect P satisfies

$$\mathrm{MSE}(Y, I|P) = (n-1)^{-1}\,\mathrm{SSE}(Y, I|P),$$

where the corresponding sum of squares $\mathrm{SSE}(I|P)$ is

$$\mathrm{SSE}(Y, I|P) = \mathrm{SSE}(Y|P) - \mathrm{SSE}(Y|PI).$$

The F statistic for the hypothesis that $\sigma_i^2 = 0$ is then

$$F(Y, I|P) = \frac{\mathrm{MSE}(I|P)}{\mathrm{MSE}(Y|PI)} = \frac{N-u}{n-1}\frac{\mathrm{MSE}(Y|P)}{\mathrm{MSE}(Y|PI)} - \frac{N-u-n+1}{n-1}.$$

Thus maximization of $\hat{\rho}_{2r}^2(Y)$ corresponds to maximization of $F(Y, I|P)$ by appropriate choice of the c_j. In terms of customary terminology in multivariate analysis (Bock, 1975, pp. 215–216), the maximum possible value of $F(Y, I|P)$ is

$$\frac{n-u-n+1}{n-1}\lambda,$$

where λ is the largest ratio $\frac{\mathrm{SSE}(Y,I|P)}{\mathrm{SSE}(Y|PI)}$ for any choice of c_j. The value of λ is customarily produced by computer procedures for multivariate analysis of variance with dependent variables X_j, $1 \leq j \leq J$, in testing whether examinee effects are present for any of the variables given that each variable satisfies an additive model in prompt and examinee. It follows that the maximum value of $\hat{\rho}_{2r}(Y)$ is

$$M_r = \frac{\lambda - (n-1)(1+\lambda)/(N-u)}{\lambda + r^{-1} - (n-1)(1+\lambda)/(N-u)}.$$

For the X_j, $1 \leq j \leq 6$, used in the construction of $\hat{C}$ in Table 29, $N = 2595$, $n = 562$, $u = 7$, and $\lambda = 3.327$. It follows that $M_1 = 0.772$, $M_2 = 0.871$, and $M_4 = 0.931$, so that there are linear combinations of essay features that yield reliable measurements of essay

Table 41.
Weights for Optimal Linear Combination of Features

Feature	Raw weight	Std. weight
Res. hol. score	0.338	0.244
Root no. chars.	0.085	0.444
Median SFI	-0.139	-0.229
Root GU	-6.703	-0.310
Root spelling	-5.518	-0.311
Root OM	-9.503	-0.471
Root Simpson's index	-6.557	-0.135

quality with only a few essays. To examine the weights c_j, it should be noted that the c_j are only determined up to a scalar multiplier. As a consequence, the listed raw weights in Table 41 apply to c_j selected so that Y has a sample standard deviation of 1. The standardized weights in the table are obtained by multiplying c_j by the root means squared error of X_j from use of a model additive in prompt and examinee. The standardized weights adjust for the variable scales of the variables X_j. These weights are the c_j that would be obtained if X_j were replaced by $(X_j - \bar{X}_j)/s(X_j)$. The optimal c_j yield a somewhat plausible measure of essay quality to the extent that the signs of coefficients are appropriate and relative magnitudes are not unusual.

An obvious question concerning these results is the extent to which users of essay scores would be comfortable with a measure of essay quality with a relatively modest weight for human scoring. It is also desirable to consider use of the methodology in this section on data from testing programs such as GRE to determine if similar results are obtained. Results from such testing programs can also clarify the role of content words in analysis of the type considered in this section. The use of $\hat{\rho}^2_{2r}(Y)$ rather than $\hat{\rho}^2_{1r}(Y)$ reflects both convenience and a desire to obtain the best possible reliability results. It is possible to perform a similar analysis for $\hat{\rho}^2_{1r}(Y)$, but somewhat more work is involved.

5 Conclusion

The data considered in this report suggest that combinations of human holistic scores and machine-generated essay features can yield considerably more reliable results than are

obtained by human holistic scoring. The extent to which such combinations can yield more desirable validity results than human holistic scores requires further investigation.

The human holistic score most favored in this study has been average holistic score, although it is certainly possible to employ the summary holistic scores testing programs provide for individual essays. It is also possible to use a single human holistic score in analysis. Resolved holistic score has been used when necessary in this report, but this form of scoring appears to be less desirable than are many alternatives.

The machine-generated essay features most emphasized in this report are generally closely related to essay features that have been previously considered. The square root of number of characters is a nonlinear transformation of a standard measure of essay length that is generated in e-rater. The median SFI is also generated by e-rater. Spelling errors have been considered in essay assessment, although the use of a nonlinear transformation of the rate per word of spelling errors does not appear to have been employed. Similarly, the square root of the rate per word of errors in grammar and usage does not appear to have been used, but e-rater has employed the rate per word of errors in grammar and the rate per word of errors in usage. The square root of the rate per word of errors in mechanics other than spelling does not appear to have been employed, but e-rater does use the rate per word of errors in mechanics. Simpson's index as a measure of word diversity does not appear to have previously been used in essay assessment, although e-rater has sought to measure word diversity. The regression analysis on square roots of rates per content word does not appear to have an obvious analogue in the literature, although both e-rater and latent semantic analysis seek to measure the distribution of content words. The analysis in this report has emphasized nonlinear transformations in order to yield essay features with distributions more similar to a normal distribution. Except for the rates of errors in grammar, usage, and mechanics other than spelling, the features used are quite easily computed.

It should be noted that changes of features in the regression in e-rater version 2.0 to features proposed in this report cannot be expected to have a major impact on quality of predictions of holistic scores. Changes may lead to more intuitive signs in regression analysis or may affect problems of outliers. In terms of current practice, replacement of

number of words and number of words squared by a transformation of number of words or number of characters may be desirable. Replacement of the types-to-tokens ratio by the square root of Simpson's index and replacement of the fifth smallest SFI by median SFI are steps likely to provide a more appropriate indication of the contributions of word diversity and SFI to the prediction of holistic score. Replacement of content vector analysis by regression with square root of word frequencies appears appropriate given the greater ease of analysis. Some minor improvements in spell checking to avoid rejection of prompt words are desirable.

Especially given recent movement toward use of essays in standard assessments, much more information on reliability is needed. External validity data are also quite important. The author did not have access to such data in preparation of this report.

References

Attali, Y., Burstein, J., & Andreyev, S. (2003). *E-rater Version 2.0: Combining writing analysis feedback with automated essay scoring.* Unpublished manuscript.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York: McGraw-Hill.

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science*, *7*(2), 96–99.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework* (College Board Report No. 99-3, GRE Board Research Report No. 96-12R). New York: College Entrance Examination Board.

Breland, H. M., & Jenkins, L. M. (1997). *English word frequency statistics: Analysis of a selected corpus of 14 millions tokens.* New York: College Entrance Examination Board.

Breland, H. M., Jones, R. J., & Jenkins, L. (1994). *The College Board vocabulary study* (ETS RR-94-25). Princeton, NJ: ETS.

Brennan, R. L. (1992). *Elements of generalizability theory.* Iowa City, IA: ACT Publications.

Burstein, J., Chodorow, M., & Leacock, C. (in press). Automated essay evaluation: The Criterion online writing service.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*, 283–284.

Cramér, H. (1946). *Mathematical methods of statistics.* Princeton, NJ: Princeton University Press.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.).

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, *63*, 435–447.

Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. Burstein, (Eds), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahway, NJ: Lawrence Erlbaum.

Finn, P. (1977). Computer-aided description of mature word choices in writing. In M. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 69–90). Urbana, IL: National Conference of Teachers of English.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.

Gini, C. (1912). *Variabilità e mutabilità: Contributo allo studio delle sistribuzioni e delle relazioni statische*. Bologna, Italy: Cuppini.

Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Erlbaum.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated readability index, fog count and flesch reading ease formula) for Navy enlisted personnel* (Research Branch Report Branch Report 8-75). Memphis, TN: Naval Air Station.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahway, NJ: Lawrence Erlbaum.

Leacock, C., & Chodorow, M. (2003). Automated grammatical error detection. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 195–207). Mahway, NJ: Lawrence Erlbaum.

Mazzeo, J., Schmitt, A., & Cook, L. (1986a). *The compatibility of adjudicated and non-adjudicated essay scores on the ATP English composition test with essay.*

Mazzeo, J., Schmitt, A., & Cook, L. (1986b). *The compatibility of adjudicated and non-adjudicated essay scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA, April 19, 1986.

Miller, R. G. (1964). A trustworthy jackknife. *Annals of Mathematical Statistics*, *35*, 1594–1605.

Page, E. B. (1967). Statistical and linguistics strategies in the computer grading of essays. In *COLING-67: Vol. 1. Conference internationale sur le traitement automatique des langues.* Retrieved May 1, 2004, from http://acl.ldc.upenn.edu/C/C67/C67-1032.pdf.

Page, E. B. (1968). Analyzing student essays by computer. *International Review of Education*, *14*, 210–225.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, *62*, 127–142.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: John Wiley.

Reid, S., & Findlay, G. (1986). WRITERS WORKBENCH analysis of holistically scored essays. *Computers and Composition*, *3*, 6–32.

Scheffé (1959). *The analysis of variance.* John Wiley.

Shermis, M., Burstein, J., & Bliss, L. B. (2004). *Can automated essay scoring impact FCAT writing scores?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA, April 13, 2004.

Simpson, E. H. (1949). The measurement of diversity. *Nature*, *163*, 688.

Slotnick, H. (1972). Toward a theory of computer essay grading. *Journal of Educational Measurement*, *9*, 253–263.

Smith, E. A., & Kincaid, P. (1970). Derivation and validation of the automated readability index for use with technical materials. *Human Factors*, *12*, 457–464.

Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York: John Wiley.